# STUDIES OF SELECTION BIAS IN APPLIED STATISTICS

By

GUNNAR EKLUND

UPPSALA 1959

ALMQVIST & WIKSELLS BOKTRYCKERI AB

From the institute of statistics at the university of Uppsala
Director: Herman Wold

# STUDIES OF SELECTION BIAS IN APPLIED STATISTICS

By

GUNNAR EKLUND

UPPSALA 1959
ALMQVIST & WIKSELLS BOKTRYCKERI AB

R. UNIV.
BIBLIOTHEEK
LEIDEN

## Contents


General Introduction . . . . . 6

Chapter 1. — *Plan and Purpose* . . . . . 9
1.1. Introduction . . . . . 9
1.2. Aspects of estimation and interpretation problems . . . . . 11
1.3. Solution approach and models . . . . . 17
1.4. Outline of the monograph . . . . . 18
1.5. Terminology, notation and presentation of solution approaches 19

Chapter 2. — *General considerations on basic hypotheses used in model construction* . . . . . 20
2.0. Introduction . . . . . 20
2.1. Some examples of different types of initial information . . . . . 20
2.2. Initial information concerning causal direction . . . . . 21
2.3. System of causal relations . . . . . 25
2.4. Independence, association and zero-correlation . . . . . 33
2.5. Factor catalogue . . . . . 34
2.6. Additivity and linearity. Interaction . . . . . 38
2.7. No effect on dispersions or moments . . . . . 41
2.8. The strength of the correlation known . . . . . 41

Chapter 3. — *Selection mechanisms. Survey* . . . . . 43
3.1. The concept "selection bias" . . . . . 43
3.2. Nonrepresentative data . . . . . 43
3.3. Partially incomplete data . . . . . 44
3.4. The occurance of M-problems . . . . . 44
3.5. Selection bias due to a common causal factor . . . . . 46
3.6. Selection mechanisms illustrated by arrow schemes . . . . . 47
3.7. Causal mechanisms and the factor M . . . . . 50
3.8. Concluding remarks . . . . . 50

Chapter 4. — *Incomplete data problems* . . . . . 51
4.0. Introduction . . . . . 51
4.1. Some examples of M-problems . . . . . 51
4.2. Causal mechanism behind the missing observations . . . . . 53
4.3. A short survey of the literature . . . . . 57

CHAPTER 5. — *New approaches relevant to incomplete data* . . . . . 60
5.1. Estimation aspects of approaches I–III . . . . . 60
5.2. Assumptions relevant to the approaches I–III. Survey . . . . . 61
5.3. Notation . . . . . 61
5.4. Approach I . . . . . 62
5.5. Approach II . . . . . 74
5.6. Approach III . . . . . 79
5.7. Approach IV . . . . . 87
5.8. Approach V . . . . . 95
CHAPTER 6. — *Selection bias due to common causal factor* . . . . . 98
6.0. Introdution . . . . . 98
6.1. Arrow scheme . . . . . 98
6.2. Investigations in which a selective factor w is suspected . . . 98
6.3. Current methods of treating this selection . . . . . 101
CHAPTER 7. — *New approaches in treating bias due to a common causal factor* . . . . . 108
7.1. Introduction . . . . . 108
7.2. General remarks on the presentation . . . . . 108
7.3. Models and approaches . . . . . 110
CHAPTER 8. — *Causal direction problems* . . . . . 131
8.0. Introduction . . . . . 131
8.1. A new approach . . . . . 131
8.2. References . . . . . 133
APPENDIX. — *Some applications* . . . . . 134
REFERENCES . . . . . 142

## Preface

Now that this work is finished I take the opportunity to express my warmest thanks to Professor *Herman Wold* for his inspiring interest in my work and for all the profit I have derived from so many valuable discussions.

My thanks are also extended to my colleagues at the Institute of Statistics, expecially to Mr. *A. Gadd*, who has read the manuscript and pointed out some omissions. Among others who at an early stage contributed imaginative suggestions I should like to mention Mr. *E. Leander* and Mr. *A. Raud*. Mrs *E. Svenonius* has translated parts of the monograph. I extend my thanks to Miss *L. Mäetalu*, Miss *M. Laadi* and Mr *P. Koskel* for valuable assistance.

Finally, I would like to express my gratitude to the *Statens Samhälls- och Rättsvetenskapliga forskningsråd* for financial assistance.

Uppsala, April 1959. *Gunnar Eklund*

# GENERAL INTRODUCTION

For a number of years I have as a consulting statistician come into contact with different research problems, principally in the fields of medicine, odontology and the social sciences.

From my experience it appeared that although the majority of statistical problems are specific—often even unique—for certain investigations, there do exist types of problems which are met repeatedly. This is especially the case, as far as I can see, for problems concerning bias through selection, among which I include bias due to incomplete data. In the present monograph I have concentrated on such questions and attempted to systematize my experiences as a consultant. In the following, I shall both discuss the currently used solutions of selection bias problems and their applicability and also present new approaches. A number of these I have already tested in my consulting activity—some will be used for illustrations; other solution approaches are as yet untried, but also for these I shall attempt to discuss their applicability.

Selection bias has been noticed among statisticians long before this monograph was planned. During the middle of the 1940's when I began my studies in statistics, I very quickly came into contact with some of these problems, which were then being discussed in the statistics seminar in Uppsala. Professor Wold, who led this seminar, had in earlier writings discussed different selection problems, especially those relevant to demand analysis and econometrics (cf. Wold, 1940). Later, in a number of publications, he developed his ideas in more detail; I refer here in particular to Wold–Juréen, *Demand Analysis* (1952) and Wold, *Causal inference from observational data* (1956). In this last-mentioned work Wold also gives a survey of ends and means in statistical investigations. Here it is proposed that selection problems occur mainly in investigations which have explanatory aims and are based on nonexperimental material, i.e. they belong to a category of investigations where, according to Wold, only sporadic and as yet unsystematized solution approaches are available—a category, therefore, which stands in need of intensive study. I might point out here that the majority of modern text books and research publications in statistics deal with experimental planning and estimation in random samples, and that experimental statistical test methods take a limited account of bias

problems and therefore cannot, without reservation, be applied to nonexperimental material.

As concerns solution approaches for selection problems, it has been emphasized in Wold's writings that a *model constructed on the basis of theoretical arguments* often leads to a solution. With the help of Pareto-Slutsky's indifference curves e.g. it is possible to derive certain relations between prices, consumption and income. This type of solution is possible only if there is a well-developed theory—and preferably no alternative theories—in the studied area. This should be the case in demand analysis, where it is often possible to rely on an economic theory. In addition there is the demand that the plausibility of the theory (model) can be verified on the basis of observation data, taken if possible from different sources, e.g. from both time series and household studies.

It has long been clear to me that a well-developed and generally accepted theory has been lacking in the areas with which I have come into contact as a consultant. In these areas it is hardly possible, on the basis of present knowledge, to construct the ingenious models required for obtaining useful results with Wold's solution approaches. I have therefore in my approaches developed a somewhat different line from Wold's, my essential purpose being to provide solutions in the area where a well-developed theory is lacking and where there are apt to be many competing explanations. Only certain simpler *models applicable in standard situations* will be considered. In this respect my approaches complement those of Wold—there is a "division of labor". Our work has gone on side by side and with mutual contacts, and of course I have drawn much inspiration from Wold's ideas.

There is a fundamental similarity in the two approaches which must be touched upon, namely our attitude towards the concept of *causality* as used in statistical applications. Wold has developed his ideas in connection with econometrics (Wold, 1952, 1954), and has later in a work for a philosophy forum (Wold, 1958) criticized the theory of causality, current among some philosophers and social scientists, which was initially put forward by Hume and has lately been expounded by Bertrand Russell. These reject the causality concept and the latter proposes to introduce in its place the concept of "predictability according to a law". In most natural sciences, however, expressions such as "cause and effect", "stimulus and response", are still used in spite of Hume's and Russell's critique.

Without going into philosophical questions, I shall follow the general opinion held by naturalists and statisticians that the use of such expressions is fruitful. For my problems and solution approaches, it is of fundamental importance to know if a certain factor can be designated as the cause of

another; and for this reason the causality concept must be made more precise.

Two types of selection bias will be treated; one can be characterized as *selection due to incomplete data*, and the other as *selection caused by a hidden factor*.

To give solution approaches which eliminate or in any case reduce this bias by holding it within certain limits is the principal aim in this monograph.

I distinguish between these two types of selection problems, both because they require *different solution approaches* and because they *occur in different types of situations*. In the presentation of all my approaches, I assume that one is given an estimation problem and certain more or less realistic initial information; I wish to show how on the basis of such information, one can construct *hypothetical models* and, using observation material, estimate different parameters referable to these models. The worth of these estimates will, of course, depend upon the realism of the model (initial information).

These two types of selections differ with respect to what is initially given; it follows that the relevant models will be different and therefore require different approaches.

As regards the situations where the problems discussed here occur, we note that bias caused by a hidden factor is found mostly in nonexperimental material where the aim of the investigation is explanatory. Bias due to incomplete data, on the other hand, can be found in experimental as well as nonexperimental materials, in explanatory investigations as well as descriptive ones.[1]

---

[1] This division into experimental versus nonexperimental, descriptive versus explanatory is systematized in the four-field table introduced by Wold (1954).

CHAPTER 1

# PLAN AND PURPOSE

## 1.1. Introduction

The purpose of this monograph is to present different approaches suitable for statistical treatment of two commonly occurring types of selection problems, namely, selection bias due to *incomplete data* (missing observations or nonresponse) or *nonrepresentative data*, and selection bias *caused by some hidden factor(s)*.

Bias due to *nonrepresentative data* has for a long time attracted great interest especially in mail and interview investigations. A well-known example from Gallup (1938) is the following (quoted from Wilson, 1952, p. 39):

"One of the most famous cases of the effect of a biased sample was the *Literary Digest* poll before the United States' presidential election of 1936. Millions of ballots were mailed out, received back, and tabulated, with the result that the easy election of Landon over Roosevelt was confidently predicted. Unfortunately the names of the mailing list were mostly taken from telephone directories, and it turned out that a strong correlation existed between wealth and a preference for Landon."

Bias introduced through *incomplete data* may be illustrated by a follow-up study of a sort often met in clinical investigations.

One wishes to investigate the results of different types of cancer operations by a follow-up study five years later. Some of the patients cannot be contacted at this later date, some for example have died from cancer and, of course, cannot be questioned. If one disregards these cases, there is risk of introducing a bias. If it had been possible to follow up all the patients, another conclusion would perhaps have been reached.

In the present monograph we have put under a common head selection arising from nonrepresentative data and selection due to incomplete data. Common to both, is that one wants to make conclusions that are valid for a certain population and has complete information for only a part of the population. In the first-mentioned example we wish to have information concerning the whole voting body—or alternatively a random sample from it—but we have data from only a biased sample. In the second example we wish to have complete statements for the whole group under investigation, but we have such statements for only the individuals contacted in the follow-up study.

In one respect, however, we have a somewhat more favorable starting position in this last example in that we know for every individual which operation has been performed. The incompleteness of the data is *partial*, for only the statements of results are concerned.

In dealing with the type of selection which is due to a *hidden common causal factor*, different approaches have been used. Selection problems of this last type appear mainly in nonexperimental material; e.g. the classical examples in Yule's famous paper (1926): Why do we sometimes get Nonsense Correlations between Time Series? As early as 1910, in his textbook, Yule gives another instructive example:

"Suppose that we have established ... a positive association between inoculation and exemption from smallpox in a universe of persons. It is natural to infer that this association is due to some causal relation between the two attributes and may be expected to recur in the future; in short, that smallpox is prevented by vaccination. This rather hasty conclusion might, however, meet an opponent who argues in this way: vaccination is accepted among the well-to-do classes, but is looked on with suspicion by the lower classes. For this and other reasons most of the unvaccinated persons are drawn from the lower classes. But these are precisely the people whom, from the unhygienic conditions under which they live, one would expect to be exposed to infection and who, moreover, being malnourished, would be more likely to contract disease when they were infected. Hence the comparative exemption of the vaccinated persons is not due to the fact that they have been vaccinated, but to the fact that they belong to the well-to-do classes."

(See Yule–Kendall, 1937, p. 50. The quotation agrees with Yule (1911) with respect to meaning but not wording.)

In this example the "social class" is clearly seen to constitute a hidden common cause factor, affecting the interpretation of an observed association between "inoculation" and "exemption".

In addition to the complex of questions which are illustrated above we shall also take up the connected problem met in ascertaining *causal direction*, i.e., whether there is a mutual interaction between two factors or whether one factor is to be interpreted as the causal (explanatory) factor and the other as the effect factor—and, in the latter case, which of the two factors is to be regarded as the cause factor. In the above example from Yule the causal direction is hardly problematic: "inoculation" can be a cause of "exemption" but not vice versa. Judging from my experience it is quite usual that the causal direction, as in this example, does not give trouble; there are, however, many exceptions. An example taken from J. P. English, F. A. Willius and J. Berkson: Tobacco and coronary disease. *JAMA*. (1940) is commented by J. R. Miner, in *Biometrics*, 1945):

"These authors found that in the age group from forty to forty-nine years the proportion of smokers among men who had coronary sclerosis was significantly greater than among those who did not have it and conversely that the incidence of coronary disease among smokers was greater than among nonsmokers. Furthermore, among smokers its incidence increased with increasing degree of smoking. In discussion of this paper Weiss pointed out that this statistical association between smoking and coronary disease did not necessarily imply causation and English and his associates "presumed that in all probability the smoking of tobacco was not an etiologic factor but was perhaps an influence that affected the course and progress of the disease."

What is the cause and what the effect is not immediately evident—if, in this case, the factors do indeed exert a real influence on each other. It is not even impossible that certain persons for the very reason that they have heart troubles are apt to smoke more.

## 1.2. Aspects of estimation and interpretation problems

The present work is concerned with the three problems outlined below; in all of them the relation between two factors (variables) $x$ and $y$ are studied. We shall deal more thoroughly with the problems in Chaps. 3–8, where the emphasis is on the bias arising from incomplete data, a common causal factor or an unclear causal direction.

The three problems to be outlined are summed up under the headings: I. explanatory *regression-problems* II. *causal direction-problems* and III. *association-problems.*

I. *Explanatory regression-problems*[1]

Background: We assume here that we know beforehand that $x$ can be the cause of $y$ but not vice versa. This situation can be illustrated in the following way where the arrow symbolizes the causal direction (cf. p. 25)

$$x \rightarrow - y$$

Two relevant questions:

(1) Does the factor $x$ have positive, negative or no effect at all on $y$? (*The Sign Problem.*)

*An example:* Is the genuine effect of inocculation ($x$) on exemption ($y$) positive or negative or nonexisting? (Yule 1911, cf. 1.1) By "genuine effect" we mean here the effect remaining after account has been taken of such disturbing factors as have effect on $x$ as well as on $y$, e.g., the social class. We denote such disturbing factors by $w_i$ and formulate the question:

[1] In these problems, as in the majority of regression analyses, we can point out beforehand an explanatory variable, $x$, and a dependent variable, $y$.

Under I are also comprised approaches related to regression analysis, such as comparisons between two groups with respect to percentage, mean, standard weighted values etc.

Is $\beta_{yx.w_1 w_2 \ldots w_n} > 0$ or $< 0$ or $= 0$ [1], where $w_1, w_2, \ldots, w_n$ is or is not observed. [2] By $x = 1$ we denote "inocculated" and by $x = 0$ "not inocculated"; by $y = 1$ we denote "exempted" and by $y = 0$ "not exempted". A disturbing factor $w$, social class, is quantified in the following way: $w = 1$ denote "worker; $w = 2$ denote "middle classes"; $w = 3$ denote "well-to-do".

(2) How strong is this effect? (*The Coefficient Problem.*) In Yule's example (cf. section 1.1) one would ask here about the value of $\beta_{yx.w_1 w_2 \ldots w_n}$ and not only about the sign as above. We can formulate the question more specifically:

What influence or effect has $x$ on $y$, after considering $w_1, w_2, \ldots, w_n$ (after $w_1, w_2, \ldots w_n$ have been *accounted for* or *eliminated*)?

*Remark:* Selection arising from incomplete data as well as that due to a "common causal factor" may be relevant.

II. *Causal direction-problems*

Background: We assume here that we know nothing of the causal *direction* beforehand.

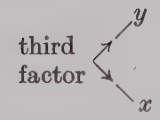
Question: We wish to know whether $x$ has an effect on $y$ or vice versa, or whether there is a mutual interaction or whether no genuine association at all is present. (*The Direction Problem.*)[3]

If, for example, we found that $x$ and $y$ were correlated we could reformulate the question: Which of the models below, A, B, C or D is applicable? (Regarding the arrow diagram cf. Wright 1934, Simon 1954.)

A $x \rightarrow - y$ i.e., $x$ has an effect on $y$ (one-way causal dependency, unilateral relation)

B $x - \leftarrow y$ i.e., $y$ has an effect on $x$ (unilateral relation)

C $x \underset{-}{\overset{\rightarrow}{\leftarrow}} y$ i.e., mutual interaction, (two-way causal dependency, bilateral relation)

D 

i.e. no genuine effect or association (cf. spurious correlation).

[1] The notation follows Yule-Kendall op. cit.

[2] This formulation, however, is limited in that $\beta_{yx} \ldots$ corresponds to the simplest linear regression model. By $\beta_{yx} \ldots$ we denote a L.S. estimate for a given population.

[3] "Genuine" correlation (or effect) is used here as the opposite of "spurious". This genuine correlation (effect) is reached after account has been taken of certain "*disturbing*" *factors*. What disturbing factors must be taken into account is discussed more thoroughly in section 1.22 and 2.5.

For illustrations, see examples e and f (1.24).

*Remark:* This complex of problems will be treated in more detail first in Chap. 8.

III. *Association-problems*

Sometimes one is inclined to disregard the causal direction and sometimes the causal direction is unknown; then one might discuss only the genuine correlation or association between $x$ and $y$.

(a) Is the correlation between $x$ and $y$, after account has been taken of $w_1$, $w_2$, ..., $w_n$, positive, negative or nonexisting? Is $\varrho_{yx.w_1w_2\ldots w_n} > 0$, $< 0$ or $= 0$? (This question concerning $\varrho_{yx.w_1\ldots w_n}$ is answered at the same time as the corresponding question about $\beta_{yx.w_1\ldots w_n}$ since the signs for $\varrho_{yx}\ldots$ and $\beta_{yx}\ldots$ will be the same.)

(b) What is the magnitude of the coefficient of correlation (strength of associations) after account has been taken of $w_1$, $w_2$, ... and $w_n$?

*Remark:* One usually asks about the correlation and association coefficient mainly in those situations where model C or D is *known* to be realistic and sometimes also if A or B is realistic. The regression-problem, on the other hand, is meaningful only if either model A or B is known to be valid.

## 1.21. *Some examples*

In order to give a more concrete picture of what is involved in these three problems—I, II, and III, respectively, we shall discuss some examples. The examples will be limited to investigations where one wishes to study the effect of one or more factors on a dependent factor; i.e., we shall limit ourselves to so-called explanatory investigations.[1]

Starting from these examples we shall, in another place, discuss certain difficulties in the analysis and interpretation of such material.

EXAMPLE a.—*Problem:* How does the consumption of liquor change with rising income? *Observation material:* Data concerning a nation's annual liquor consumption and average income for every year within a certain time period. (Cf. Malmquist, 1948.)

EXAMPLE b.—*Problem:* Is mortality due to lung cancer dependent on smoking? *Observation material:* Data concerning mortality due to lung cancer within a group consisting of heavy smokers and the corresponding mortality within a group of nonsmokers. (Cf. Berkson, 1955.)

[1] Investigations which attempt to give explanations and which are based on non-experimental data occur very frequently within many of the sciences, above all in medical and the social sciences. Here belong several so-called clinical investigations, econometric applications, sociological field investigations, criminological and demographic investigations, etc.

EXAMPLE c.—*Problem:* Do dentures made of acrylic give rise to a lower frequency of palatal ulcers (denture sore mouth) than dentures of rubber or vice versa? *Observation material:* Data concerning a selected group of denture wearers. Observations have been made as to whether or not these people are subject to palatal ulcer and whether the material of their dentures is of rubber or acrylic. (Cf. Nyquist, 1952.)

### 1.22. *Commentary on the examples*

These examples, although they come from different fields of research, show certain similarities:

(i) In no case has the observation material been produced through experiment.

(ii) An attempt is made to determine one factor's effect upon another:

In example a, one asks: What effect does a rise in income (change of income) have on liquor consumption?

In example b, one asks: What effect does smoking have on lung cancer?

In example c, one asks: What effect does denture material have on palatal ulcer?

The first-named factor (variable) is designated as *explanatory*, denoted $x$, the latter as *dependent*, denoted $y$.

(iii) In all the examples it is possible, on the basis of the observation material, to estimate the correlation coefficient between the explanatory and the dependent variable in order to clarify the problem—tentatively in any case.

(iv) There are interpretation difficulties throughout. The correlation coefficients will be very difficult to interpret; it is quite doubtful if they really give relevant answers to the respective questions.

If the correlation in example a turns out to be negative, several explanations could be given: The correlation might be negative because increased incomes lead to lower liquor consumption, but the more probable explanation is that rising prices are accompanied by increased incomes and it is really the rising prices which have brought about a negative correlation and therefore not the rising incomes (i.e. spurious correlation between income and consumption). Similar interpretation difficulties may also be found in the other examples. For instance in example c, it should be noted that older dentures in a majority of the cases are of rubber, while recently acquired dentures are made of the more modern acrylic material. Data indicating higher palatal-ulcer frequency among wearers of rubber dentures can be explained by the fact that the dentures have been used a longer time or that the wearers are older, but cannot alone be proof that acrylic dentures are better. The price factor in example a, the age of the dentures and that

of the denture wearers in example b, we designate as selective factors, denoted $w$.

(v) The questions are vaguely formulated and need to be clarified. The interpretation difficulties are in part due to the vagueness of the questions. In example a, for instance, one asks: What factors must be considered? In estimating the association between income and liquor consumption, it is necessary to take account of the price level, as mentioned above. Must we also consider the number of the adult population, the age distribution, the number of cars etc.?[1]

We shall return to this problem in section 2.5. There we shall also discuss in detail the method of establishing the factor catalogue which is relevant to a given problem.

Three further examples follow below; example d involves an experimental situation and examples e and f concern situations where a bilateral causal influence is suspected.

### 1.23. *The concepts factor and variable*

Example d illustrates (1) the regression problem, where causal direction is initially known, and (2) the concept *factor* as opposed to the concept *variable*.

Let us consider two factors: the fluoride concentration (in the form of salts) of drinking water and the development of caries. (For reference concerning fluoride as a means against tooth decay, see *Svensk Tandläkaretidskrift* 1954).

Concerning these two factors we know at once that there can be only a unilateral causal connection; fluoride concentration may have some effect on caries but not vice versa.

If the fluoride concentration is denoted $x$ and caries $y$, the causal direction is then represented by model $a$ (see 1.2), i.e. $x \rightarrow y$.

Relevant questions:

1. What effect does the fluoride concentration of drinking water have on the development of caries? Possible answers are that increased fluoride concentration increases, decreases or does not affect caries activity.

2. (If some effect is present) How great an effect does the fluoride concentration have on caries?

The last question cannot, of course, be dealt with unless every factor is quantified, i.e. replaced by a variable. For the "fluoride concentration" some of the following variables may be introduced: (i) fluoride concentration per liter of water, (ii) fluoride

---

[1] Is the immediate effect of the income rise intended or its long term effect? And are income rises per individual or for a whole population to be considered? The average income can remain unchanged, but at the same time an income rise can occur through an increase in the population number.

weight in mg consumed per day, (iii) more than $x_0$ mg per liter would correspond to $x = 1$, less than $x_0$ mg to $x = 0$.

A measure for "caries activity" might be: (i) changes in the number of teeth intact during the observation period (the more intact, the less caries), (ii) changes in the number of front teeth intact, (iii) changes in the number of decayed surfaces.

In this monograph the word "factor" will be used in a somewhat wider meaning than "variable". We use the word "variable" when quantities (measurements) are introduced. By "factor" we intend a qualitative concept. A given factor can be quantified in many ways, i.e. can be represented by variables differently defined, as appears in the above example.

The type of problem which we have in example d can be illustrated by experimental as well as nonexperimental material. Nonexperimental studies comparing places with high and low fluoride concentration were later followed by experimental studies in which the water was artificially fluoridized. In the nonexperimental comparisons, it is possible that the differences with respect to caries frequency are completely or in part due to substances other than fluoride salts, e.g. such substances as occur in great quantities in ground water; places with high fluoride concentration use ground water directly, while places with low concentration—usually large cities—have often had to resort to artificial water. In the fluoridizing experiments this source of error can be avoided.

This underlines a well-known fact: In experimental material those interpretation difficulties which arise from a common causal factor can be completely avoided (cf. Wold, 1952, p. 31).

### 1.24. *Some examples of unclear causal mechanism (the direction problem)*

Example e.—This example illustrates the situation $x \rightleftarrows y$. We start from an almost generally accepted argument (cf. Welander, 1955): The development of caries ($x$) depends among other things on retention sites. Now, a pocket formation ($y$) implies a new retention site; this can in its turn be thought to lead to a caries lesion in the environment of the pocket formation. It can also be suspected that caries in the vicinity of the alveolar region tend to give rise to gingivitis which in its turn leads to pocket formation. For certain individuals: caries→– pocket formation, while for others: pocket formation→–caries. Dealing then with a group of individuals observed in *cross-sectional investigations*, we must note that an observed association between caries and pocket formation in the aggregated material can be explained either by saying that caries give rise to pocket formation or that pocket formation gives rise to caries.

Example f.—This example illustrates a situation where it is unclear which of the causal direction models A, B, C or D is applicable.

A question which has attracted great interest among sociologists and psychologists is the following: Does alcoholism give rise to marital conflicts or vice versa; or are there common factors behind both these factors which give rise to the positive correlation between them? (Cf. Gösta Carlson, 1954, p. 548.)

Different explanations could be given here:

(i) alcoholism is *per se* a cause of marital conflicts. If so action taken against alcoholism should be of help to the marriage.

(ii) the conflicts in marriage are the cause of alcoholism. If so a possible therapy would be an attempt to influence the married couple's sexual and erotic relations, by way of information.

(iii) for certain families (i) holds and for others (ii). For the aggregated material we get no unique causal relation. The therapy must be varied from family to family.

(iv) the correlation is misleading and represents only an indirect causal connection; there is a common factor behind these factors, e.g. people who are inclined toward alcoholism are also apt to come into conflict. The therapy would perhaps be more complicated.

## 1.3. Solution approach and models

We shall in this monograph schematically divide every solution approach into the following steps: (1) Formulation of the problem, (2) model construction, (3) observational material (fictive or factual), (4) mathematical derivation of solution, (5) discussion, in application to concrete material, of the model's plausibility, and interpretation of the numerical results.

We have in the foregoing section touched upon the first point. After the problem is made more precise (cf. 2.53), it is usual to construct a model, often even before observation material is available. In the construction of models different types of information are used; the approach may be based on standard scientific procedures or be motivated by "common sense". The model can as a rule be characterized as hypothetical since the initial (basic) knowledge does not necessarily describe the facts. As a consequence, the estimates and conclusions which may be obtained on the basis of the observation material will be fully valid only under the condition that the model is realistic; i.e. the hypotheses are correctly specified.

It is therefore of great importance to be able to verify that the model is applicable (i.e. without a specification bias); cf. steps (2) and (5) in the schema above.

In the following chapter we shall discuss some basic considerations relating to model construction and also present in more detail different elements used in the construction.

### 1.31. *Example of a solution approach*

In order to illustrate a solution approach we shall discuss a very simple example connected with Yule's example (cf. 1.1).

1. *Problem:* Determine the sign of $r_{yx.w}$, where $x$ represents inoculation, $y$ exemption and $w$ social status (cf. association problem, III above, section 1.2). Population: The individuals included in the observation material.

2. *Model:* We assume that (i) inocculation takes place most often "among the well-to-do classes", so that $r_{xw} > 0$, i.e. there is a positive association between social class and inocculation frequency; (ii) "the lower classes" run a greater risk of falling ill, so that $r_{yw.x} > 0$, i.e. there is a positive association between "social class" and "exemption" from reasons other than vaccination (vaccination having been eliminated as a factor). Both these inequalities constitute our hypothetical initial information of the subject matter. The assumptions are seen to be plausible.

3. *Observation material:* $x$ and $y$ are observed for a number of individuals, $w$ is not observed for any.

From the concrete observation material we shall get a numerical value for $r_{yx}$. We find that $r_{yx} > 0$ or alternatively $\leqslant 0$.

4. *Mathematical deduction:* Assumptions: $r_{xw} > 0$ and $r_{yw.x} > 0$. If $r_{yx} \leqslant 0$, we can prove, as below, that $r_{yx.w} < 0$, but if $r_{yx} > 0$ this cannot be proved. We assume:

$$r_{yw.x} = \frac{r_{yw} - r_{xw}\, r_{yx}}{\sqrt{(1 - r_{xw}^2)\,(1 - r_{yx}^2)}} > 0,$$

hence

$$r_{yw} = r_{xw}\, r_{yx} + \varepsilon, \quad \text{where } \varepsilon > 0.$$

Then it follows, by substituting $r_{yw}$ as above,

$$r_{yx.w} = \frac{r_{yx} - r_{yw}\, r_{xw}}{\sqrt{(1 - r_{xw}^2)\,(1 - r_{yw}^2)}} = \frac{r_{yx} - r_{xw}\, r_{yx}\, r_{xw} - \varepsilon\, r_{xw}}{\sqrt{(1 - r_{xw}^2)\,(1 - r_{yw}^2)}} = \frac{r_{yx}\,(1 - r_{xw}^2) - \varepsilon\, r_{xw}}{\sqrt{(1 - r_{xw}^2)\,(1 - r_{yw}^2)}} < 0.$$

This inequality holds if $r_{yx} \leqslant 0$ and $r_{xw} > 0$.

As is seen, we can determine the sign of $r_{yx.w}$ only if $r_{yx}$ is negative, in other words the available initial information gives definite knowledge about the sign only under certain favorable—and in this case improbable—conditions on the observed correlation $r_{yx}$.

## 1.4. Outline of the monograph

We observe that certain initial information is necessary as a complement to the observed $r_{yx}$ if the sign problem is to be answered. The elements of this model, namely $r_{xw} > 0$ and $r_{yw.x} > 0$, constitute only some of the many types worthy of attention. We shall return to a fuller account below (Chap. 2).

In the very simple example above we have illustrated that initial information can be used in the estimation of the genuine association between two factors $x$ and $y$ (or the effect of $x$ on $y$) when a hidden selective factor,

$w$, might be expected to give rise to a bias (so that $r_{yx} \neq r_{yx.w}$). We shall, in Chaps. 6–7 return to this question.

In Chaps. 4–5 we shall show in more detail that initial information is significant also in the case where incomplete data affect the dependent variable $y$, so that observation material for $y$ exists for only a part of the relevant investigation population. Finally in Chap. 8 the question of unclear causal direction is treated. In Chaps. 3–8 we will illustrate how, on the basis of the model, one can derive estimators via mathematical operations.

## 1.5. Terminology, notation and presentation of solution approaches

*General remarks*

Use of terminology and notation has been summarized in 2.34, in 5.3 and in 7.2. In general the notation used in regression and correlation analysis is like that in Yule (1907). Other terminology follows Cramér (1945), however with rather many exceptions.

Finally some remarks on the presentation of my solution approaches:

(i) The emphasis in this monograph is upon the treatment of systematic error (bias) and for pedagogical reasons I shall almost completely disregard stochastic (random) error. I have therefore chosen as a rule to work with very large samples and whole populations.

(ii) From a mathematical point of view many of the models and approaches presented here are rather uncomplicated. I have not attempted to give general mathematical solutions, as I am aware that even for rather simple problems, it is difficult to construct realistic models capable of solution. Nor have I systematized the solutions into a general theory from which they could be derived as different special cases.

(iii) I have often gone so far in simplification as to limit the discussion of approaches to those cases where the studied variables are dichotomous. This has often been done simply to illustrate the basic idea of the solution approach and, as a rule, it is not impossible to apply the same idea even to continuous variables.

CHAPTER 2

# GENERAL CONSIDERATIONS ON BASIC HYPOTHESES USED IN MODEL CONSTRUCTION

## 2.0. Introduction

The purpose of this chapter is to present some important types of initial information (hypotheses) used in model construction and, in passing, to touch upon the problem of judging their plausibility (realism). Not until subsequent chapters, however, will we discuss the central question as to how on the basis of the given initial information one can obtain solution approaches effective in the treatment of selection problems. The types of initial information treated in this section have been chosen with this question in mind. The present chapter has been written because no systematic compilation of the types of initial information central to this monograph can be found in the literature.

## 2.1. Some examples of different types of initial information

Initial information in statistical applications is useful in the planning of investigations as well as in the analysis (interpretation) of material already collected. In the foregoing chapter we showed by means of an example (1.31) how certain initial information concerning correlation coefficients can be of use. Below other examples follow taken from widely differing areas, where for a long time one has been aware of the usefulness of initial information concerning (i) causal direction, (ii) independence, (iii) relevant factors, (iv) structure of the association (mathematical form), and (v) interaction. We shall later refer to this grouping.

(i) In a randomized experiment it is evident that the stimulus (treatment), $x$, can have an effect on the response (dependent variable), $y$, but not vice versa. The *causal direction* is known.

(ii) In such an experiment "underlying factors", $w$, are neutralized, and *uncorrelated with* or even *independent of* $x$.

(iii) It is desirable to have information about all *relevant factors*. In the classical "laws of nature" this is often possible; e.g. in Gay-Lussac's law, concerning ideal gases, $p.\ v/T = c$, where $c$ is a constant characteristic of a certain gas. When estimating the constant, all factors are known.

(iv) Gay-Lussac's law can be sharpened. The behaviour of ideal gases is better approximated by the mathematical *function* known as van der Waal's equation:

$$\frac{(p + a/v^2)(v - b)}{T} = c;$$

$a$, $b$ and $c$ are constants, characteristic of a certain gas. This equation can be derived from the kinetic gas theory and can be regarded as a theroretical model.

(v) In planning experiments one often assumes that *interactions* (of high order) can be omitted. This assumption is used especially in "confounded design" (Fisher, 1935; Yates, 1933).

In the following we shall first discuss questions concerning the definition of causality (2.2) and then present more complicated systems of causal relations (2.3). Using the arrow scheme, we shall attempt to systematize initial information of the types found under (i), (ii), and (iii) above (causality, independence and relevant factors). Initial information concerning independence is treated in (2.4), that concerning relevant factors (factor catalogue) in (2.5). That about interaction, additivity and linearity is discussed in (2.6) and in connection with this, the more general question of the structure of the association (or its mathematical form) will be considered. Initial information concerning dispersion is treated in (2.7). In (2.8) we shall deal with situations where we have information about the strength or sign of the correlation.

## 2.2. Initial information concerning causal direction

### 2.21. *The concept of causality*

Causality problems are central to what follows in two respects:

(1) In both the interpretation and practical consequences of the investigation results—the interpretation and the operation aspects. (Cf. example f: If we can state that alcoholism is a cause of marital conflicts, this is a more specific and, from a practical point of view, a more important conclusion than simply to state that a correlation exists.)

(2) In solution approaches in Chaps. 3–8 where causality constitutes an important element in many models.

For the discussion which follows it is necessary to make clear: (i) the meaning which is given to causality and causal hypothesis (2.22)— the definition aspect; (ii) the situations where the causal direction is evident (2.23)—the realism aspect.

### 2.22. *The definition aspect*

The concept of causality has been an object of discussion for a very long time. It is not my purpose to restate the arguments; from a philosophic-

statistical point of view they have been commented on by Wold (1958) to whom reference is made. As concerns the meaning of causality, I follow Wold's definition (1952, 1954, 1958) where causality is defined as a hypothetical element in a model. Wold (1958) begins by defining "causal relation" in controlled experiments:

"The stimulus variables of the experiment are called cause variables, or simply causes; the response variables are called effect variables, or effects. The relation between cause and effect variables that enters in the hypothesis connected with a controlled experiment is called a causal relation."

Then he proceeds to nonexperimental situations:

"Having now covered the case of controlled experiments, we proceed to nonexperimental situations. The definition of causal concepts extends to this case by way of a fictive parallel: Three groups of variables $x$, $y$, $z$ are called cause variables, effect variables, and disturbance variables, respectively, if the theoretical model includes the hypothesis of a fictive controlled experiment with the $x$'s for cause variables, the $y$'s for effect variables, and the $z$'s for disturbance variables."

### 2.23. *The realism aspect*

The above definition is clearly independent of the question whether it is realistic to presuppose that a causal influence is present in a given concrete situation. In principle there is no reason why we could not set up a completely unrealistic causal model. The practical conclusions that could be drawn from such models are, however, of very limited or no worth. From a standpoint of operation, the model must be realistic if the conclusions are to be usable. We shall therefore in the following sections (to section 2.3 inclusive) explain by means of examples the situations where we hold it justified to speak of *realistic causal hypotheses*.

The realism problem can be of highly varying degrees of difficulty. Examples e and f in the foregoing, e.g., illustrate situations where it is clearly unrealistic to proceed on the assumption of a simple-directioned causal influence. However in the experimental approach mentioned in example d (fluoride concentration→−caries), there is no reason to doubt that the causal direction is unique.

The realism of the causal direction is evident above all in experimental situations.[1] By experimenting with varying doses of fluoride (explanatory

[1] Cf. the following comment, Wold & Juréen, 1952: "Since experiments are not possible, the causal interpretation is not obvious from the mechanism of an experiment but has instead to be justified by considerations ad hoc. Generally speaking, the causal interpretation has to be inferred from subject-matter considerations based on empirical evidence and economic theory."

variable) one can show that a stronger dose leads to a decrease in caries activity (dependent variable).

The following examples show that the causal direction is also evident in many nonexperimental situations. The argumentation here must be based on subject-matter consideration:

(1) A question naturally asked in the medical field is whether (increased) cigarette smoking or some other factor is the cause (explanation) of the rise in lung cancer frequency observed in U.S.A. and in western Europe. (Cf. example b.)

(2) It is natural for economists to suppose that an increase in the price of liquor leads to (causes) a decrease in liquor consumption.

(3) Among psychologists a question long discussed is whether genetic factors can explain (are the cause of) differences between individuals with respect to intelligence (or other qualities).

The above examples show that in many fields a causal point of view is generally applied.[1] It seems that such a viewpoint has been accepted by the majority of those working in a great number of sciences where statistical applications are relevant.

### 2.24. *Some examples*

We shall, beginning with some examples, show that realism problems in experimental and nonexperimental situations are rather related, and also, using these examples, make more precise what is here meant by causality according to the definition.

*An example from criminology:* It is well known that men commit considerably more serious crimes than women. Should this be interpreted as a cause–effect relation, i.e. sex →– criminality?

*An example from medicine:* Blindness is more usual among aged than among young persons. Can one then say that old age is a cause of (a high frequency of) blindness?

*An example from caries research:* Through experiment it has been found that children whose teeth were treated with sodium fluoride application showed lower caries activity than did an untreated group. What can be inferred about the causal connection?

*Comments on the examples:*

(1) Sex (age and sodium fluoride, respectively) cannot possibly be assumed to be the only causal factor underlying criminality (blindness, caries activity, respectively). Sex (age, sodium fluoride) can be neither a necessary nor a sufficient cause.

(2) An inverted causal direction is impossible (unrealistic): criminality cannot affect sex; blindness cannot be the cause of (affect) old age; caries activity in children cannot be the reason that fluoride is applied just to certain children's teeth and not to others'. Only

[1] This often happens unreflectively; such a point of view seems self-evident in spite of the philosophers' critique.

one causal direction is clearly possible (cf. p. 12): sex→ – criminality; age→ –blindness; sodium fluoride→ – caries activity. Nor is there a question of mutual interaction here; a bilateral relation is precluded.

(3) In none of these examples the causal effect is immediate (*direct*). E.g., with growing age, body tissues, including those in the eyes, degenerate. Certain changes of this sort lead to impaired sight. The effect of growing age is indirect, it works via other factors. Most probably there is a causal chain: Old age→ – tissue change→ – impaired seeing ability.

From an etiological point of view, old age could hardly be called a cause of impaired seeing ability. In the present monograph, however, these indirect relations will also be described as causal.

From the above examples it may be seen that we assume that sex can affect criminality, old age can affect blindness, etc. Our causal concept implies: (i) a causal factor is not equivalent to "a single causal factor"; (ii) it is possible that a causal factor may be neither a "*sufficient* cause" nor a "*necessary cause*"; (iii) a cause may be—and as a rule is—indirect, i.e. causal factors work via one or more intermediate factors in a causal chain.

### 2.25. *Causal direction and correlation*

In the earlier discussed example from Yule the causal direction is evident (inoculation $(x) \rightarrow -$ exemption $(y)$). A positive correlation, $r_{xy} > 0$, is not in itself any proof that vaccination is effective; one might well suppose that the positive correlation is due to a third factor, social class.

### 2.26. *Causal direction versus effect*

The fact that only a single causal direction is possible does not necessarily imply the existence of an effect worth mentioning.

If $CaCO_3$ were applied instead of fluoride, undeniably the only possible causal direction would be $CaCO_3 \rightarrow -$ caries activity. It would, however, be quite surprising if some noticeable effect on caries activity were present. It is possible to find situations where the causal direction of a certain factor is clear enough, but where its effect is unclear, i.e. is positive (favourable) for certain individuals and negative (unfavourable) for others. Certain drugs, amphetamine, caffeine etc. have a stimulating effect on certain people and an anti-stimulating effect on others. But in spite of this the causal direction is clear.

### 2.27. *Summary*

In the present work we follow the definition of causality given by Wold & Juréen (1952) and Wold (1958). This definition, however, has no bearing

on the realism problem. Questions concerning the realism of a causal model have therefore been illustrated through several examples and will be further developed in section 2.3 where systems of causal relations will be treated.

## 2.3. System of causal relations

### 2.31. *General remarks*

In order to describe graphically the causal relations between different factors it is natural to connect them with an arrow in the following way:

Causal factor $\rightarrow$ – result factor (effect).

This method of diagramming is found in the statistical literature already in Sewall Wright (1918) and also in Wright's papers on the so-called "path analysis",[1] or the method of path coefficients (1921, 1934) and in Tinbergen (1939). Every path here corresponds to an arrow.

We have used the arrow diagram earlier in this monograph to describe simplified situations where there is only a single causal relation. We shall, in the following, illustrate and comment on more advanced systems, namely those where the connection between many factors is studied; we shall attempt to give a realistic picture of the "flow of causation" in these systems.

The causal systems to be described here, however, are *hypothetical*. This does not alter the fact that more *realistic* systems have decided advantages. We shall return to certain aspects of this question in section 2.5.

### 2.32. *Some central concepts illustrated by an example*

The following example taken from a study of pigs (see Wright, 1934, and Tukey, 1954) illustrates a causal system comprising three variables (factors).

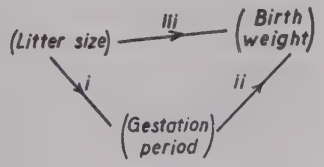



Fig. 1.

[1] Wright's papers have not attracted great attention. Not until the symposium *Statistics and Mathematics in Biology* was published in 1954 did his ideas become generally known, both through his own article and through Tukey's article in the same publication. For a later presentation of Wright's ideas, see Kempthorne (1957) and Li (1956).

The three arrows correspond to the following causal relations: (i) Litter size presumably affects the length of the gestation period—while the converse relation would seem to be precluded. (ii) The gestation period in its turn can be assumed to have an effect on birth weight (estimated per litter for every maternal parent). (iii) Litter size affects the birth weight, but not vice versa.

We can imagine that the litter-size effect goes in two different paths: one via gestation period, the other indirectly, via unspecified factors (i.e. the lower and the upper paths in Fig. 1). Why just these two paths should be of interest is apparent from Wright's initial question: Why is there a negative correlation between birth weight and litter size? Minot (1891) proposed alternative paths of competition for limited nourishment and stimulation of premature birth (see Tukey, 1954, p. 60).

We shall extend and more fully comment on this example of a causal system, without however concerning ourselves with Minot's particular questions.

1. By bringing together these two paths—the upper and the lower—we get the following *reduced arrow scheme:*

$$(\text{litter size}) \rightarrow - (\text{birth weight}).$$

This arrow scheme is relevant, e.g., if one wished to know the average effect of litter size on birth weight. This, however, is a different problem from the one originally posed.

2. *The observation material* (factual or fictive) in Wright's example should preferably have consisted of data on each of the maternal parents concerning her litter size, gestation period and the litter's birth weight.

3. If one wishes to introduce an *additional factor* in this causal system, namely the age of the maternal parents—a possible cause of all the factors mentioned—the arrow scheme could be constructed in the following way. However again we do not claim that the scheme is complete or realistic:

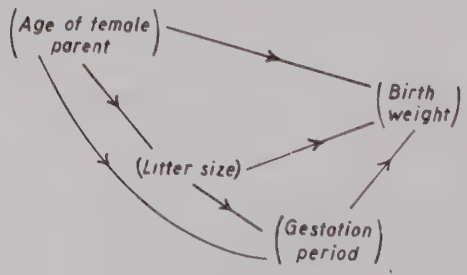


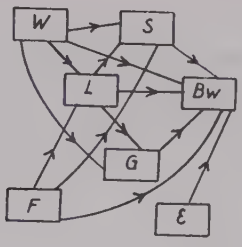

Fig. 2.

4. Knowing only the causal direction, we cannot tell whether the sign of the association is positive or negative. E.g. larger litters would perhaps give rise to shorter gestation periods (negative association) but this cannot be seen from the arrow scheme.

5. Fig. 1 illustrates an *incomplete* causal system. If we try to determine all the factors required to explain variations in the birth weight of litters taken from different pigs, it is not sufficient to consider only litter size, maternal parent's age and gestation period. For a realistic model we should also consider factors such as sex distribution among the pigs, maternal parent's qualities, time of year (season) of birth, maternal parent's diet during gestation period etc.

We can set up a rather comprehensive *catalogue of factors* thought to have an effect on birth weight. In the same way we can set up a factor catalogue for gestation period and litter size. And similarly for the sex distribution among pigs, maternal parent's qualities, time of year of birth, and maternal parent's diet etc.

Let us write all the above named factors as direct or indirect causes affecting birth weight, gestation period and litter size and then introduce an arrow scheme. We shall obtain a set of relations and there will be many branchings in the arrow scheme (see Fig. 3). And again this is certainly incomplete.




*Bw:* Birth weight
*L* : Litter size
*G* : Gestation period
*S* : Sex distribution
*F* : Season of birth
*W* : Age of mother, diet etc.
*ε* : Weighting error.

Fig. 3.

6. There is hardly reason to assume that there is a real association between certain of these factors. It is difficult to see how observation error (see Fig. 3), could affect sex distribution or vice versa. Observation error seems to be *independent* of all factors. In Wright's terminology observation error is a *residual factor*. (Observe that no arrow has observation error as a "target".)

7. Nor is there an arrow connecting sex distribution and gestation period; this is due to the fact that sex distribution has no direct effect on the length of the gestation period or vice versa. It cannot be denied, however, that there may be a correlation between sex distribution and gestation

period. But the reason this correlation occurs is that other common factors have an effect on sex distribution as well as on gestation period—litter size and the age of the maternal parent could be such factors.

8. Two types of causal systems can be distinguished:

(i) those which concern one individual,

(ii) those which concern a group of individuals (e.g. a population).

The causal systems which are given above can be classified as both (i) and (ii). In the latter case maternal parents constitute the individuals in the group. Example (e) p. 16 underlines the difference between the two types of systems.

i) For certain individuals it holds: caries→–pocket formation  For others it holds: caries–←pocket formation;

ii) For a group of individuals it holds:

$$\text{caries} \rightleftarrows \text{pocket formation}.$$

The grouping of causal relations is obtained by superposing the individual systems one on another (aggregation).

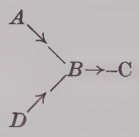
9. The last named example illustrates a bilateral association occurring through aggregation. However, it is often the case that causal systems, without being less realistic, are free from such bilateral relations (this clearly holds in Figs. 1–3).

### 2.33. *On patterns of causal influences—some illustrations*

1. *Causal chain:* $A \rightarrow - B \rightarrow - C$. Simon (1954) has described $B$ as an *intervening* factor in a situation where the association between $A$ and $C$ is studied.

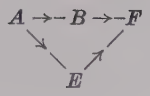
2. *Reduced chain:* If the intervening factor, $B$, is left out the chain reduces to: $A \rightarrow - C$.

3. *Independence:* $A$ and $D$ are independent. Both $A$ and $D$ have an effect on $C$ (indirectly).



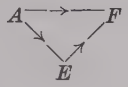

4a. *Partial independence:* In the causal chain $A \rightarrow - B \rightarrow - C$, $A$ and $C$ are independent, after the effect of $B$ has been taken into account. $A$ has an effect on $C$ only via $B$. If the model is correctly specified, the partial correlation is $\varrho_{AC.B} = 0$; i.e. no partial correlation exists between $A$ and $C$ after $B$ has been considered.

4 b.




Here $A$ and $F$ are independent, after the effects of $E$ and $B$ have been taken into account (cf. $\varrho_{AF.BE} = 0$).

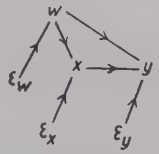
If B is left out we obtain:




If both $E$ and $B$ are left out, we get: $A \rightarrow - F$.

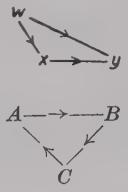
5. *Underlying factors:* In the arrow scheme factors in the beginning of a causal chain are placed to the left and the dependent factors to the right. In the above arrow schemes $A$ is always on the left and is said to be a factor underlying $B$, $C$, $E$ and $F$. Analogically $D$ underlies $C$, while $B$ underlies $C$ and $F$.

6. *Residual factors:* The residual factors are denoted by $\varepsilon$ with subscripts corresponding to the relevant target.




A residual factor is never a target. The arrow scheme is often reduced by disregarding the residual factors.

7. *Causal circle:*




A causal circle can describe the result of an aggregation, e.g. an aggregation over subjects (cf. p. 28) or an aggregation over time periods. The latter aggregation can be derived from a causal chain:

$$A_t \rightarrow - B_t \rightarrow - C_t \rightarrow - A_{t+1} \rightarrow - B_{t+1} \rightarrow - C_{t+1} \rightarrow - A_{t+2} \ldots,$$

The following relations are assumed to hold:

$$A' = A_t + A_{t+1}, \quad A'' = A_{t+2} + A_{t+3}, \quad \text{etc.}$$
$$B' = B_t + B_{t+1}, \quad B'' = B_{t+2} + B_{t+3}, \quad \text{etc.}$$
$$C' = C_t + C_{t+1}, \quad C'' = C_{t+2} + C_{t+3}, \quad \text{etc.}$$

where $A'$, $A''$, $B'$, $B''$, $C'$, $C''$ etc. are observed, but every $A_t$, $B_t$, $C_t$ etc. is not observed.

Let $A$ denote $A'$, $A''$, $A'''$ etc. ($B$ and $C$ analogically), then the arrow scheme for $A$, $B$ and $C$ can clearly be described as a causal circle. A causal circle is reduced into a bilateral causal system if $C$ is left out.

$$A \overset{\rightarrow}{\underset{\leftarrow}{-}} B$$

### 2.34. *Survey and notation*

$A \rightarrow\!\!-\, B$ Read: The factor $A$ has or can be thought to have an effect on $B$, but not vice versa. This must be true for every subject in the population. The effect can be indirect or direct (with or without an intervening factor).

$\underset{?}{\rightarrow\!\!-}$ Read: This effect is studied.

$x \underset{?}{\rightarrow\!\!-}\, y$ Read: We will estimate a coefficient (or the sign) for the effect of $x$ on $y$.

$A \overset{\rightarrow}{\underset{\leftarrow}{-}} B$ Read: The factor $A$ has an effect on $B$ for some subjects and $B$ has an effect on $A$ for other subjects.

$A$——$B$ Read: The relation between $A$ and $B$ may be indirect or direct, unilateral or bilateral. The correlation between $A$ and $B$ may also be spurious, due to a common factor (cf. p 12).

$x$ = explanatory factor (variable), the effect of which is studied; $y$ = dependent factor (variable); $w$ = one common causal factor or a complex of such factors, underlying both $x$ and $y$. We denote different values of the variable $w$ by $w'$, $w''$, $w'''$ etc., and different values of $x$ by $x'$, $x''$, $x'''$ etc. By $w_1$, $w_2$, $w_3$ we mean three different common causal factors.

### 2.35. *Quantification of arrows*

One aim in our causal system presentation is to characterize one or more arrows by a coefficient. In Fig. 3 we ask: How great is the effect of $L$ on $Bw$, if the effects of other factors in the figure are taken into account. The coefficient for the effect may of course vary between sub-populations; for that reason we concentrate our interest on the mean effect within the whole population under consideration. We shall illustrate how such a quantification is carried out by Sewall Wright.

## 2.36. "*Path analysis*"

Sewall Wright, in *AMS* 1934, has given a survey of The method of path coefficients. Some essential aspects of this method are presented in the following quotation:

"The method of path coefficients was suggested a number of years ago (Wright 1918, more fully 1920, 1921), as a flexible means of relating the correlation coefficients between variables in a multiple system to the functional relations among them. The method has been applied in quite a variety of cases.

Those variables which are treated as dependent are connected with those of which they are considered functions by arrows. The system of factors back of each variable may be made formelly complete by the introduction of symbols representative of total residual determination (as $V_u$ in Fig. 1). A residual correlation between variables is represented by a double-headed arrow. It will be assumed that all relations are linear." (Wright, 1934, p. 161.)

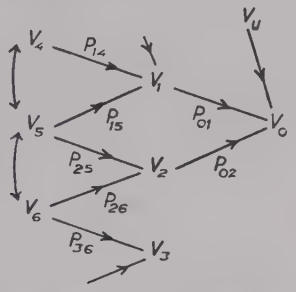



FIG. 4. (From Wright, 1934, Fig. 1. I have added the path coefficients, $P_{ij}$.)

*Comments*

(i) The residual correlation between $V_4$ and $V_5$ can be the result of a common underlying factor. Wright denotes residual correlation by a double-headed arrow.

(ii) $V_4$ and $V_6$ are not connected by a double-headed arrow, i.e. $V_4$ and $V_6$ are not correlated.

(iii) Wright adds a supplementary assumption concerning linearity; this cannot be read from the arrow scheme itself.

(iv) Every path coefficient $P_{ij}$, where i stands for the dependent and $j$ for the explaining variable, is connected with an arrow. $P_{ij}$ is a type of partial regression coefficient corresponding to a situation, where every variable is standardized with respect to its standard deviation.

In Fig. 4 we could thus express the relation between the dependent variable $V_0$ and the explanatory variables:

$$\frac{V_0-\overline{V}_0}{\sigma_0}=P_{01}\frac{V_1-\overline{V}_1}{\sigma_1}+P_{02}\frac{V_2-\overline{V}_2}{\sigma_2}+P_{03}\frac{V_3-\overline{V}_3}{\sigma_3}+P_{0u}\frac{V_u-\overline{V}_u}{\sigma_u}.$$

(The notation is as usual. $V_u$ is a residual factor. We know beforehand that $P_{03} = 0$, because no path connects $V_3$ and $V_0$.)

There is a rather simple relation between path coefficients and regression coefficients:

$$P_{01} = \frac{\sigma_1}{\sigma_0}\beta_{01.23u}, \quad P_{02} = \frac{\sigma_2}{\sigma_0}\beta_{02.13u} \quad \text{etc.}$$

The corresponding regression coefficients are found in the following expression:

$$V_0 - \overline{V}_0 = \beta_{01.23u}(V_1 - \overline{V}_1) + \beta_{02.13u}(V_2 - \overline{V}_2) + \beta_{03.12u}(V_3 - \overline{V}_3) + \\ + \beta_{0u.123}(V_u - \overline{V}_u).$$

Tukey (1954) has emphasized that the regression coefficients—being more stable—are preferable to the path coefficients.

If the linear regression approach is applicable, a regression coefficient corresponds to every arrow.

Path coefficients are related also to correlation coefficients. Wright shows e.g. that $r_{0g} = \sum P_{0i} r_{gi}$.

If all factors are taken into consideration, we obtain the identity-correlation:

$$r_{00} = \Sigma P_{0i} r_{0i} = 1.$$

Wright further remarks (op. cit., p. 163):

"That two factors affect the same dependent variable does not contribute to the correlation between them. Similarly two variables which are correlated with a third are not necessarily correlated with each other. As illustrations of these principles consider the correlations between some of the variables in Fig. 1.

$$r_{36} = P_{36} \quad r_{46} = 0 \quad r_{13} = P_{15} \cdot r_{56} \cdot P_{36}$$

$$r_{23} = P_{23} \cdot r_{56} \cdot P_{36} + P_{26} \cdot P_{36} \quad r_{12} = P_{14} \cdot r_{45} \cdot P_{25} + P_{15} \cdot P_{25} + P_{15} \cdot r_{56} \cdot P_{26}.$$

In estimating the path coefficients Wright starts from information concerning correlation coefficients, which are computed by means of observations. As seen above simple relations exist between values of $P$ and $r$. He thus obtains an equation for every numerical $r$ and by using the relations, numerical values for $P$ can be derived from a system of equations.

* * *

In the preceding section, 2.3, the fundamental concepts of independence, factor catalogue and linearity were used in constructing causal systems and in estimating parameters corresponding to arrows. The following sections,

2.4–2.6, are relevant to the model constructions in Chaps. 3–8. The material is compiled from different authors; for further references and a more detailed treatment I refer to Yule (1911), Ezekiel (1941), Wold (1952), Tukey (1954) and Wright (1954).

## 2.4. Independence, association and zero-correlation

### 2.40. *Definition*

The concept "association" is used with somewhat varying meanings. In the following we enumerate three possible meanings of the expression: $y$ and $w$ are not associated.

(i) The conditional distribution of $y$ relative to every $w$ must be equal: $\phi(y|w') = \phi(y|w'')$, where $w'$ and $w''$ are possible values of $w$ and $\phi(y|w)$ stands for cumulative frequency distribution. This is the sense in which we will use the concept "independent" in this monograph.

(ii) The following conditional means must be equal:

$$E(y|w') = E(y|w'').$$

If definition (i) is valid it follows that (ii) also is valid, but not vice versa.

(iii) The correlation coefficient equals zero; i.e. $\varrho_{yw} = 0$. If (i) or (ii) is valid, it follows that $\varrho_{yw} = 0$, but not vice versa.

It is possible to generalize the definitions in (i)–(iii). For instance we may add a new variable $x$; (i), (ii), and (iii) are then reformulated:

(i) $\phi(y|w', x^{(k)}) = \phi(y|w'', x^{(k)})$ holds for every fixed value of $x$.
(ii) $E(y|w', x^{(k)}) = E(y|w'', x^{(k)})$ holds for every fixed value of $x$.
(iii) $\varrho_{yw.x} = 0$.

The above assumptions concerning $\phi$, $E$ and $\varrho$ refer to *population* distributions, expectations and correlations respectively.

### 2.41. *Correlation, regression, and arrow scheme*

If no association exists according to (ii) above, the factors $y$ and $w$ are not connected with an arrow.

In Fig. 5 the variables $z$, $w$, $x$, and $y$ are included. If the linear regression approach is applicable, a regression coefficient corresponds to every arrow (see Fig. 5).

If $z$ and $w$ are independent Fig. 5 can be reduced (the arrow between $z$ and $w$ can be eliminated as in Fig. 6):

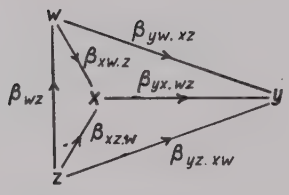



Fig. 5.

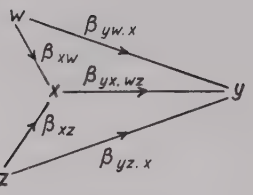



Fig. 6.

If the true relation between the factors were *linear*, a zero correlation, $\varrho_{zw} = 0$, would be sufficient in order to eliminate the arrow.

## 2.5. Factor catalogue

By "factor catalogue" we mean a list of factors having direct or indirect effect on a variable. One factor catalogue can be set up for the dependent variable, $y$, and another for the independent variable $x$ etc..

### 2.51. *Reduced factor catalogue when estimating a regression coefficient*

If, however, we want to study the effect of $x$ on $y$ (the regression problem, see 1.2), it is possible to construct a *reduced factor catalogue*.

If $\beta_{yx.u} = \beta_{yx}$, the factor $u$ can be eliminated from the factor catalogue.

In the following three situations $u$ can be eliminated (cf. *Wold & Juréen*, 1952, Ch. 12.3.2):

(i) If $u$ is correlated with $y$ but not with $x$, and if the relation between $y$, $x$, and $u$ is linear, then $\varrho_{yu} \neq 0$ and $\varrho_{xu} = 0$. From $\varrho_{xu} = 0$ it follows that $\beta_{ux} = 0$, and since $\beta_{yx.u} = (\beta_{yx} - \beta_{yu}\beta_{ux})/(1 - \beta_{ux}\beta_{xu})$ (cf. *Yule-Kendall*, 1937, p. 269) we obtain $\beta_{yx.u} = \beta_{yx}$.[1]

(ii) If $u$ has no other effect on $y$ than via $x$, i.e. if $\beta_{yu.x} = 0$, then $\beta_{yx.u} = \beta_{yx}$. To this may be added:

(iii) If $u$ is an *intervening* factor in relation to $x$ and $y$, the factor $u$ should not be included in the factor catalogue.

---

[1] Note that as a rule $\varrho_{yx,u} \neq \varrho_{yx}$, and that the signs of the correlations are the same, since

$$\varrho_{yx,u} = \frac{\varrho_{yx} - \varrho_{yu}\,\varrho_{xu}}{\sqrt{(1-\varrho_{yu}^2)\,(1-\varrho_{xu}^2)}} = \frac{\varrho_{yx}}{\sqrt{1-\varrho_{yu}^2}}.$$

Hence if $\varrho_{yu} \neq 0$, it follows that $\varrho_{yx,u} \neq \varrho_{yx}$.

### 2.52. *Difficulties in constructing the factor catalogue*

That all factors included in the factor catalogue should be considered, is in principle agreed upon by statisticians (cf. *Ezekiel*, 1941, Chap. 23), but in practice it is almost never possible to find all relevant factors and it is difficult to obtain observation data for all factors.[1]

In section 2.3 we discussed how the causal system is changed by introducing more factors. The following example is another illustration.

### 2.53. *Illustration of factor catalogue and arrow scheme. Specification of the estimation problem*

Let us state the problem: What effect does a rise in income have on liquor consumption? The observations consist of cross-sectional individual data. (Cf. example a.)

We assume hypothetically that our factor catalogue is complete. Since this assumption is a hypothesis, the practical conclusions drawn from estimations based on the material depend upon the realism of the assumptions.

*Notation:* By $x$ we denote change in income over a period of two years, by $y$ change in liquor consumption, by $u = 1$ buying a car, and by $u = 0$ not buying a car.

A rise in income is often followed by buying a car. The new car owners probably decrease their liquor consumption (due to the punishment for drunken driving). Further, by $v$ we denote change in saving, by $t$ change in consumption objects other than cars, by $\varepsilon_u$, $\varepsilon_x$, $\varepsilon_t$, $\varepsilon_v$, $\varepsilon_y$ we denote residual

---

[1] Cf. the following quotation from *Wold-Juréen*, 1952, p. 36:

"Causal factors of secondary interest must in nonexperimental data be taken as they come, and their possible disturbing effect cannot be neutralized by the randomization design of controlled experiments. Such factors will be referred to as *secondary factors*. The traditional device in this situation is to include the secondary factors among the regressors. In other words, to obtain a regression relation in which the regression coefficients give a correct picture of the causal influence of the factors of primary interest, the set of regressors is enlarged by the secondary factors. The main dilemma of this method is that the secondary factors usually are numerous, whereas in practice the regression analysis must be confined to a limited number of regressors. The principle traditionally followed to meet this difficulty is to include those variables as regressors that according to experience and a priori theory are believed to be the main causal factors of the regressand."

(Cf. also *Wold*, 1956.) All the "secondary factors" mentioned by Wold should be included in the factor catalogue.

factors not affecting factors other than the factor in the subscript, and by $w$ a complex of factors such as sex, education, overtime work and so on; these factors perhaps have an effect on both $y$ and $x$.

*Population:* Adult inhabitants in Sweden.

*Arrow scheme:*

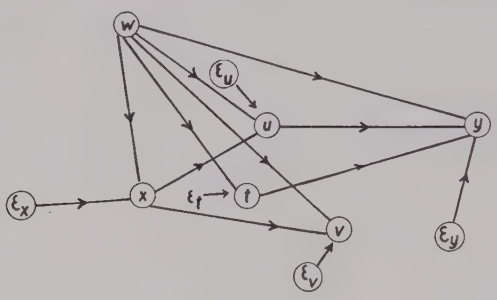



Fig. 7.

In order to determine which factors should be included it is necessary to specify the problem.

*Specification of the problem:*

Some different ways of specifying problems are possible. Each specification has a characteristic factor cataloque.

$\alpha$) *Experimental ceteris paribus situations.* We specify that all factors affecting personal income as well as consumption of liquor must be considered (controlled). Such a factor is e.g. overtime work which increases the income and perhaps limits the opportunities of liquor consumption.

If now an increase in income were given at random to some individuals (the experiment group) but other individuals (the control group) were not given an increase in income, a difference in the liquor consumption between the two groups could be explained by the factor "increased income" but not by any underlying factor (if random errors can be ignored). This example illustrates how a nonexperimental problem can be reformulated as if an experiment had been undertaken.

$\beta$) *Partial ceteris paribus situations.* In addition to the same factors as those in ($\alpha$), even certain factors affecting liquor consumption (and perhaps affected by income increase) should be considered. "Buying a car or not", $u$, is such a factor. One can then ask: What effect on liquor consumption does an increase of income have within a sub-population consisting of individuals who have not bought a car? (Alternatively: have bought a car.)

$\gamma$) *The extereme ceteris paribus situations* (usually meaningless). All factors affecting income are considered. For instance prices, saving, consumption of foods, clothes etc. are all controlled factors. Considering these factors we want to study the effect of an increase in income.

Problems ($\alpha$), ($\beta$) and ($\gamma$) illustrate situations where different regression coefficients are relevant. In ($\alpha$) the aim is to estimate the effect of $x$ on $y$; we will estimate $\beta_{yx.w}$. No intervening factors, $u$, $v$, $t$ should be considered in the analysis. In ($\beta$) the aim is to estimate $x$-effect on $y$, after taking account of $u$ and $w$, i.e. $\beta_{yx.wu}$. In ($\gamma$) we try to estimate $\beta_{yx.wtuv}$. According to the diagram, $\beta_{yx.wtuv} = 0$; $x$ has an effect on $y$ via $t$, $u$ and $v$. These arguments emphasize the point that not all the factors affecting $y$ must be included. The inclusion of intervening factors has an important effect on the interpretation. Not until the problem is specified is it possible to decide which, if any, of the intervening factors should be included.

*Factor catalogue for y. Experimental "ceteris paribus" situation.* Complete factor catalogue for $y$:

$$t,\ u,\ v,\ x,\ w,\ \varepsilon_u,\ \varepsilon_v,\ \varepsilon_t,\ \varepsilon_x,\ \varepsilon_y.$$

The factor catalogue relevant to the problem can be reduced: $\varepsilon_y$ can be eliminated because $\varrho_{x\varepsilon_y} = 0$, $\varepsilon_u$ because $\varrho_{x\varepsilon_u} = 0$, $\varepsilon_v$ because $\varrho_{x\varepsilon_v} = 0$, $\varepsilon_t$ because $\varrho_{x\varepsilon_t} = 0$, and $\varepsilon_x$ because $\varrho_{y\varepsilon_x.x} = 0$ (cf. 2.51).

The factors $t$, $u$, and $v$ can be eliminated as they are not relevant to the problem. The factor $w$ must, however, be included, since $w$ has a (direct) effect on both $x$ and $y$.

## 2.54. *Factor catalogue and regression analysis. Bias due to an incomplete factor catalogue*

For a population we want to estimate a parameter, e.g. $\beta_{yx.z}$. If we had observed $y$, $x$ and $z$-values for all subjects in the population or in a random sample, we could by means of the least-squares method, estimate $b_{yx.z}$. For this estimator we know that $E(b_{yx.z}) = \beta_{yx.z}$[1] (disregarding the magnitude $0(1/n)$). In this case our factor catalogue fits our problem perfectly.

In the more general case we want to estimate $\beta_{yx.zw_1w_2\ldots w_k}$ for a population, where $z$, $w_1$, $w_2$, ..., $w_k$ denote factors that are correlated with both $x$ and $y$, i.e. $\varrho_{yw_i.xzw_1w_2\ldots w_kw_j\cdots w_k} \neq 0$ and $\varrho_{xw_i.zw_1w_2\ldots w_kw_j\ldots w_k} \neq 0$ is true.

Further we assume that all $w_i$ have the same position in the arrow scheme as did $z$ in the preceding example, i.e. $w$ has an effect on both $x$ and $y$.

Generally the inequality $\beta_{yx.z} \neq \beta_{yx.zw_1w_2\ldots w_k}$ is valid. Note that it is pos-

[1] If the whole population is observed $b_{yx.z} \equiv \beta_{yx.z}$.

sible to obtain a complete catalogue only if we have observed ($y$, $x$, $z$, $w_1 \dots w_k$) for every subject.

As a rule we have to accept the fact that the factor catalogue is incomplete in one respect or other, e.g. that we know none of the factors $w_1 \dots w_k$. This situation is characterized as selection due to a "common causal factor". If we now try to estimate the $\beta$-value on the basis of the observation-data—data concerning $y$, $x$, and $z$—using the least-squares estimator $b_{yx.z}$ (or $b_{yx}$), we would find that this estimator is biased.

### 2.55. *The most important factors*

As a rule it is impossible to find a complete catalogue but it can still be argued that a more complete catalogue is better than a less complete one. Therefore we should consider (observe) such $w_i$-factors for which $\varrho_{yw_i \cdot xzw_1 w_2 \dots w_h w_j \dots w_k}$ as well as $\varrho_{xw_i \cdot zw_1 w_2 \dots w_h w_j \dots w_k}$ (or the product between them) has a great numerical value.

## 2.6. Additivity and linearity. Interaction

### 2.61. *Concept of linearity*

Three types of assumptions concerning linear models can be specified: (1) valid for every subject, (2) valid for subjects in a sub-population only, (3) valid only as an adjustment (fitting of a line).

(1) We assume that the genuine relation for the subject (i) is

$$y = \alpha + \beta x + \varepsilon_i, \text{ where } E(\varepsilon_i) = 0;$$

i.e. for every subject for which $x = x'$, the expected value of $y$ is: $E(y|x') = \alpha + \beta x'$.

(2) We do not assume that $E(\varepsilon_i) = 0$ is valid. We assume, however, that the mean of $\varepsilon$ is zero for the subjects in the sub-population group where $x = x'$; i.e. $E(\bar{\varepsilon}|x') = 0$. Hence for this group or for one individual sampled at random $E(y|x') = \alpha + \beta x'$, and in general it holds that $E(y|x) = \alpha + \beta x$.

(3) In this case $\beta$ refers to a linear adjustment of all the population data by fitting of a line.

### 2.62. *Reasons for using linear approaches*

Among statisticians the linear approach is common, e.g. multiple regression analysis.[1]

[1] In analysis of variance linear models are often simplified into additive ones. In "confluence analysis" it is assumed that an exact linear relation exists between $x$ and $y$. In factor analysis, discriminant analysis, and multivariate methods, linear models are used extensively (cf. Kendall, 1948).

There are many reasons for using linear approaches:

(1) If an interval is short an expansion according to Taylor's theorem of even a non-linear function $y = F(x, z)$ gives:

$$y = F(\bar{x}, \bar{z}) + (x - \bar{x})\, F'_x(\bar{x}, \bar{z}) + (z - \bar{z})\, F'_z(\bar{x}, \bar{z}) + \text{Residual term.}$$

If we disregard the residual term, we have a linear expression in $x$ and $z$ (cf. Wright 1954, p. 20).

(2) The linear approach also gives a rough approximation in the case where the interval is large.

(3) The linear approach is often sufficiently realistic. (Cf. 2.61 (1) and (2).)

(4) The computations are simplified considerably.

### 2.63. *Additivity*

Let us consider the expression

$$y = a + f(x) + g(z) + \varepsilon.$$

Note there is no interaction terms such as, e.g. $xz$, $xz^2$, $x^2z$ etc. Note that $x$ and $z$ are not necessarily linear. The model is said to be additive.

### 2.64. *Interaction*

Let us assume that $x$ and $z$ are dichotomous, having the values 0 or 1. We can then write $y = f(x, z)$ as $y = a + bx + cx + dxz$.

It is evident that this model contains an interaction term $xz$. Each possible combination of $x$- and $z$-values has a corresponding $y$-value. If $x = 0$ and $z = 1$, $y$ must be $a + c$. If $\varepsilon$ is added to the equation, we obtain the new expression:

$$y = a + bx + cz + dxz + \varepsilon, \text{ where } E\varepsilon = 0.$$

We now obtain

$$E(y|x = 0, z = 1) = a + c \text{ and } E(y|x = 1, z = 1) = a + b + c + d.$$

We can estimate $a$, $b$, $c$, and $d$ by multiple regression analysis, taking $y$ as a dependent variable, and $x$, $z$, and $xz$ as independent variables. What is said under 2.61 (1)–(3) applies also to interactions.

If $x$ and $z$ have no interaction-effect on $y$, it is evident that $d = 0$. We could express the situation where $d \neq 0$ by the following arrow scheme. Between $x$ and $xz$ and between $z$ and $xz$ there is a (non-causal) relationship.

Fig. 8.

*Remark 1:* Interaction terms can sometimes be eliminated by a transformation. If the expression, $p = T^{\alpha}/v^{\beta} \cdot c$, (cf. Gay-Lussac's law of ideal gases) is transformed into $\log p = \alpha \log T - \beta \log v + \log c$, there is no longer an interaction between the new variables $\log p$, $\log T$ and $\log v$.

*Remark 2:* The model mentioned in 2.64 is often used even when $x$ and $z$ are not dichotomous.

### 2.65. *Interactions and independent events*

A special type of interaction exists if the multiplicative relation valid for independent events is applicable, i.e. if $P(H_1 \cdot H_2) = P(H_1)P(H_2)$.

Given the dichotomous independent variables $y$, $x$, $z$, etc. which can assume the values 0 and 1, it follows:

$$P(y = 1 \mid x, z) = 1 - P(y = 0 \mid x, z) = (1 - b_x x)(1 - b_z z),$$

where $b_x$ and $b_z$ are constants.

### 2.66. *Limited interaction*

It is sometimes realistic to assume that the sign of the $x$-effect is the same for *all z-values*. In the figure we have assumed that $y = f(x, z)$ and that $f(x = 1, z) - f(x = 0, z) > 0$ for all $z$-values.[1]

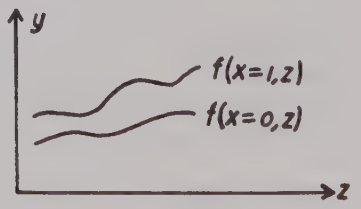



Fig. 9.

This assumption is always valid if no interaction $(xz)$ exists. The interaction may exist, however, but is of limited magnitude.

We could reformulate the above assumption in mathematical terms: The inequality

$$\{f(x = 1, z^{(i)}) - f(x = 0, z^{(i)})\}\{f(x = 1, z^{(j)}) - f(x = 0, z^{(j)})\} \geqslant 0$$

is valid for all $z$, or alternatively

$$\operatorname{sign}\{f(x = 1, z^{(i)}) - f(x = 0, z^{(i)})\} = \operatorname{sign}\{f(x = 1, z^{(j)}) - f(x = 0, z^{(j)})\}.$$

[1] Alternatively we could formulate the assumption:

$$E(y \mid x = 1, z) - E(y \mid x = 0, z) \geqslant 0.$$

If the condition above is imposed on the model in 2.64 we obtain (setting $z^{(i)}$ equal to 0 and $z^{(j)}$ equal to 1):

$$\{(a+b)-(a)\}\{(a+b+c+d)-(a+c)\} \geqslant 0$$

i.e. $b \cdot (b+d) \geqslant 0$.

This is a boundary condition for the variation of $b$ and $d$.

### 2.67. *Relations between interactions of high and low order*

If no primary effect exists, it is sometimes plausible to assume that interactions do not exist or are unimportant. If no first-order interaction exists we may sometimes assume that interactions of higher orders can be left out etc.

Assumptions of this sort are often made in statistical experimental design (see Kendall 1948, Chap. 23).

## 2.7. No effect on dispersions, moments etc.

A special type of assumption: The standard deviation of $y$ does not change even if $x$ has some effect on the mean of $y$, i.e.

$$\sigma(y|x = x^{(i)}) = \sigma(y|x = x^{(j)})$$

This assumption is common in statistical tests of significance and in calculations of confidence limits (e.g. Student's test, analysis of variance, regression analysis).

The assumption seems to be plausible especially if $x$ has no effect on $y$.

## 2.8. The strength of the correlation is known

(i) It is sometimes possible to state beforehand that $|\varrho_{xw}| < k < 1$, where $k$ is an initially known constant. This type of assumption is often plausible if we can specify the non-observed $w$-factor. Sometimes it is then possible to approximate roughly $\varrho_{xw}$ by means of a correlation $r_{xw}$ calculated in earlier investigations.

(ii) From earlier investigations or by theoretical reasoning, it may often be assumed that the correlation is positive (or negative). E.g. the correlation between

$$\begin{array}{ccc} \text{infection} & \rightarrow - & \text{falling ill} \\ w & & x \end{array}$$

is probably positive.

In section 1.31, we discussed the assumption that $r_{yw.x} > 0$ and $r_{wx} > 0$.

(iii) We can assume we know that one correlation is larger than another, e.g.

$$\varrho_{xy} > \varrho_{uw}.$$

*An example:* The association between hereditary attributes of a father ($x$) and son ($y$) ought to be stronger than the association of such attributes between grandfather ($u$) and grandson ($w$). In the same way the correlation between brothers should be greater than that between cousins. Cf. Cochran (1954).

CHAPTER 3

# SOME FUNDAMENTAL SELECTION MECHANISMS. SURVEY

The purpose of the present section is (i) to clarify the concept of selection, (ii) to describe different types of selection mechanisms and (iii) to represent selection mechanisms in the form of arrow schemes.

## 3.1. The concept "selection bias"

In discussions and interpretations of statistical analyses one often meets the expression "selection bias"; the expression has, however, a number of different meanings. Generally it is used to signify a bias which may be thought to arise from

1) nonrepresentative data (as a basis for inferences from a sample to a universe or population),

2) partially missing or incomplete observation data for the dependent variable,

3) a common causal factor not taken into account.

## 3.2. Nonrepresentative data

Selection in the first meaning—i.e. selection bias in an inference from a nonrepresentative sample to a population—is what is intended in the following excerpt taken from *Bradford-Hill*, 1956:

"In medical statistics we are nearly always working with samples drawn from large populations. For instance, if we compare the mortality of diabetics treated with insulin with the mortality of diabetics not treated with insulin, we do not possess information regarding all such persons (whom we may term "universe" of diabetics) but only for a sample of each type which has come under our own observation or under the observation of others. If we wish to argue from this sample to the universe from which it was drawn, to deduce that what is true of our sample of patients is therefore true of the general run of patients, then we must consider very carefully, whether in fact our sample is fully representative of all patients, and not in any way biased or 'selected'."

The word "selection" is used in this meaning especially in descriptive sample investigations. Another example belonging to the same type is the one taken from *Gallup* mentioned in section 1.1. Here certain individual groups are over-represented and others under-represented in proportion to the population (= voting body). We also find this type of selection in explanatory (comparative) investigations (see the example in section 3.61, $\beta$ and $\gamma$). In such investigations we define data as *nonrepresentative* if for certain individuals in a population, data are completely missing for both the independent and the dependent variable.

### 3.3. Partially incomplete data

In explanatory investigations we define data as *partially incomplete* (missing) if for certain individuals in a population we have data for the independent variable but are lacking data for the dependent variable. With regard to solution approaches this situation, in contrast to the one where data are completely missing, offers greater possibilities. This is why it is treated in itself. Strictly speaking, selection bias due to missing data is present only if the missing data frequency depends upon (or is correlated with) the $y$ variable. An example of this is found in the follow-up study in section 1.1.

### 3.4. The occurance of *M*-problems [1]

*M*-problems in general are surprisingly little represented in the statistics literature. An exception, however, are *M*-problems appearing in the form of non-response, and treated in connection with the sampling of persons questioned in interviews and those filling out questionnaires in mail investigations. (Cf. *Dalenius*, 1955, 1957 and *Hansen-Hurwitz-Madow*, 1953.) Also the most interesting solution approaches come from the field of mail and interview investigations.

*M*-problems occuring in these investigations can most often be characterized as non-representative data problems; i.e. for the case of non-response, we lack data for both the $x$ and the $y$ variable. However, when the sampling is based on certain classifications, we sometimes do have information about the $x$ variable for all individuals. For instance, social-class classifications can be found in community registers; if we want to study the influence of social class ($x$) on, say, attitudes ($y$), the incompleteness of the data would be only partial, complete for $x$ and incomplete for $y$.

[1] We introduce the abbreviation *M-problems* to denote problems arising when: (i) data for both $x$ and $y$ are missing (nonrepresentative data), (ii) data are missing only for $y$ (partially missing data).

### 3.41. *Incomplete data in follow-up investigations*

In many areas $M$-problems are only peripherally noted.

(i) This is true of clinical investigations where the patients are called in for examination. The examples from *Nyquist*, on p. 54, and *Berkson*, p. 55 are exceptions since in the majority of such investigations these problems are not taken into account.

(ii) Special problems arise in follow-up studies where the same individuals are observed on more than one occasion. A rather detailed discussion of such problems is found, e.g. in *Strindberg* (1956).

In follow-up studies, there is often data for the $x$ variable, i.e. the incompleteness is only partial.

How $M$-problems occur in follow-up investigations can be illustrated by the following statement from *Chapin* (1947, p. 167):

"Some cases disappear during the interval of time that the experiment runs. Some are lost by refusals to give information on second interview, some have become deceased, and some have moved away and cannot be found. As a consequence the terminal groups are not composed of the same individuals as at the outset. This means that at the terminal date the groups have lost their random character. Certain selective influences have been at work."

### 3.42. *M-problems in experimental investigations*

$M$-problems arise in experimental as well as in non-experimental investigations. For instance, in experimenting with animals it sometimes happens that some of the animals die. In other experiments, observations are purposely rejected because the values obtained seem to be absurd. (*Gumbel* 1958, *Wilson* 1952, pp. 256–58.) Also there is very little in the literature concerning $M$-problems as they occur in experimental investigations. The so-called missing-plot technique has been developed primarily with a view to facilitating analyses in certain experiments.[1]

In this technique the missing data are not seen in connection with the value of the dependent variable—the incompleteness of the data does not affect the construction of the model. Hence, the missing-plot technique may be described as a numerical method in variance analysis which is suitable if the possible biasing effect of incomplete data can be disregarded.

---

[1] The technique has been used for a long time and was theoretically treated in connection with variance analysis by *Yates* 1938; for an account see *Kendall* 1948, pp. 229–233.

## 3.5. Selection bias due to a common causal factor

With respect to selection bias due to a common causal factor we may distinguish two cases, treated in 3.51 and 3.52 respectively, depending on whether a regression problem or a correlation problem is relevant. The bias which is caused by a common causal factor depends on the problem which is treated.

### 3.51. *Selection bias when estimating the effect of x on y*

By a "selective-biasing" factor $w$ we mean a factor which has an effect on $x$ as well as on $y$. Selection due to such a factor often occurs when one wants to investigate the effect of one measurable factor $x$ on another measurable factor $y$ and is limited to non-experimental material. The presence of a third factor $w$ (or possibly several such factors) not taken into account can give rise to a nonsense-correlation between $x$ and $y$ if $w$ has an effect on both $x$ and $y$. The typical selection-bias situation is present if $w$ is not a measurable factor or is completely disregarded.

The interpretation difficulties (cf. Chap. 1) which confront us here have been recognised for a long time. (See e.g. *Mayo-Smith*, 1896, p. 276, regarding the relation of occupation to criminality; in this case, factors such as age, sex et. al., produce a distortion effect since they cannot be eliminated with the help of available statistics.)

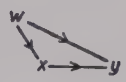
Factors mentioned in some of the preceding examples are arranged in the following table. ($x$ and $y$ are studied and observed variables; $w$ is not observed.) We want to estimate $\beta_{yx.w}$.

| Example | Explanatory variable $x$ | Dependent variable $y$ | Suspected selective factor $w$ |
|---|---|---|---|
| Yule (see p. 10) | vaccination | exemption | social class |
| a (see 1.2) | price of alcohol | consumption of alcohol | income |
| b (see 1.2) | smoking | cancer | urbanization, age, sex et. al. |
| c (see 1.2) | denture material | palatal ulcer | age of the dentures, age of the patients |
| Mayo-Smith (see above) | occupation | criminality | age or sex |

In this monograph, the factor $w$ is designated as selective only under certain conditions. In 1.2.I (where we want to estimate $\beta_{yx.w}$) it is necessary,

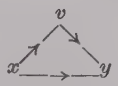
in order to correctly designate $w$ as a selective factor, that $w$ has a real effect on $x$ as well as on $y$, i.e. $w$ must be a common factor lying behind these other two factors.

*Arrow scheme:*




If the factor $w$ either has no effect on $x$ or no effect on $y$, it would hold that $\beta_{yx.w} = \beta_{yx}$.

*Remark:* If the arrow scheme looked as follows




$v$ would not be a common causal factor since $x$ has an effect on $v$, but not vice versa. In our terminology $v$ is not a selective but an intervening factor (cf. 2.33).

### 3.52. *Selection bias when estimating correlation coefficients*

In problem 1.2.III we want to investigate the genuine correlation between $x$ and $y$, i.e. the correlation remaining after $w$ has been eliminated. Since a hidden factor $w$ may be thought to have an effect on $x$ as well as on $y$ one can suspect a spurious correlation.

For this problem the factor $w$ will be designated as "selective" if it has an effect on $x$ or on $y$ or on both $x$ and $y$.

Observe that if e.g. $\varrho_{yw} = 0$ holds, $\varrho_{yx.w}$ is not as a rule equal to $\varrho_{yx}$ (cf. 2.51).

## 3.6. Selection mechanisms illustrated by arrow schemes

Using arrow schemes, we shall illustrate some of the above mentioned selection mechanisms. We limit ourselves to situations where we want to study the effect of $x$ on $y$. In addition to the earlier notation (cf. 2.34) we introduce the following:

Every person can be characterized as:

$M = 1$ not belonging to the observation sample or not wholly accounted for with respect to $x$ or $y$ (missing, non-included),

$M = 0$ belonging to the observation sample (included).

By $M$ we mean the dichotomous variable: included or non-included. Similarly, to every letter there corresponds a factor having a value which varies from individual to individual. For instance, $w$ can mean social

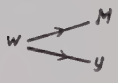
class; then $w = 3$ will stand for the well-to-do class, $w = 2$ for the middle class, $w = 1$ for the laboring class etc. The expression $w \rightarrow -M$ shall be read: social class is relevant in determining whether an individual (unit) is included or is not included in the material (sample).

The following arrow schemes concern conditions within a population and correspond to the situation represented by an aggregation of arrows (cf. p. 28). Only single-direction causal systems are relevant here.

The selection concepts with which this monograph is concerned (see above) can be illustrated by the following standard situations (type-cases): 3.61–3.63.

### 3.61. *Nonrepresentative data*

($\alpha$) *Aim:* Estimation of the percentage pro Landon within the whole voting body. (Cf. the example from Gallup) *Units:* Individuals




Variables: $y = 1$ for Landon
$y = 0$ against Landon
$w = 3$ well-to-do class
$w = 2$ middle class
$w = 1$ laboring class
$M = 1$ not included in the material
$M = 0$ included in the material.

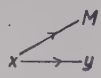
($\beta$) *Aim:* Estimation of a difference between percentages (averages) for a population. (Cf. the example from Bradford-Hill). *Units:* Individuals having diabetes.

Variables: $y = 1$ deceased during the observation period
$y = 0$ non-deceased during the observation period
$x = 1$ treated with insulin
$x = 0$ not treated with insulin
$w = 1$ non-severe cases
$w = 0$ severe cases.

Arrows are drawn from $w$ because severe cases of diabetes (in contrast to the non-severe cases) (1) are more likely to be included in the sample; (2) have a greater mortality risk; (3) to a greater extent are treated with insulin. An arrow is drawn from $x$ to $M$ because those treated with insulin certainly are more likely to be included in the observation material — such cases are usually under a doctor's treatment. The reason for the arrow from $y$ to $M$ is that due to autopsy records, persons deceased during the observation period are more likely to be discovered (therefore included in the material) than non-deceased persons.

If we find that for one of two groups being compared, a proportionally greater amount of data is missing than for the other, we usually speak of *differential selection* (cf. Wallis-Roberts, 1956, p. 302).

($\gamma$) *Aim:* (Cf. the example in $\beta$.) *Units*: Individuals having severe diabetes (for which $w = 0$)



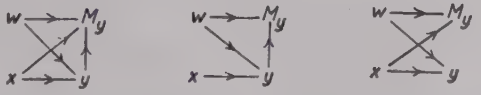

Notation as in $\beta$.
Observe that the factor $w$ is left out.

### 3.62. *Selection due to partially missing observation data*

Observation data for $y$ are lacking for certain individuals; data for $x$ are known for the whole population. $M$, therefore, has been written with the subscript $y$. Common to all partially-missing-data situations is that the factor $M$ is affected:

$\alpha$) $M$ is affected by both $x$ and $y$
$\beta$) $M$ is affected by $y$
$\gamma$) $M$ is affected by $x$
$\delta$) $M$ is affected neither by $x$ or $y$ but by $w$.

More complicated situations arise if we superpose scheme $\alpha$ on $\delta$, $\beta$ on $\delta$ and $\gamma$ on $\delta$.




### 3.63. *Selection due to a common causal factor*

### 3.64. *Commentary*

If we compare the different types of selection mechanisms described in the above arrow schemes, we find that they show evident similarities.

It is of special interest to compare the arrow schemes in 3.62 $\alpha$ with the arrow scheme in 3.63. These are alike in so far as they are all in the form of a triangle.

However an essential difference is that the arrow of most interest is in different positions in 3.62 $\alpha$ and in 3.63; i.e. the arrow between $x$ and $y$.

In 3.62 $\alpha$ the factor $M$ is a target for arrows from $x$ and $y$—in 3.63, however, $y$ is a target for arrows from $x$ and $w$.

### 3.7. Causal mechanisms and the factor $M$

Observe that $M$ is a factor like $x$, $y$ and $w$, but that it occupies a special position due to the fact that it tells us something about the structure of the observation material. For those individuals where $M_y = 1$, data are missing for $y$; where $M = 1$, data are missing for both $x$ and $y$.

From the arrow scheme it may be seen that $M$ occurs only as a dependent variable. One should not, however, be led by the above arrow schemes to believe that $M$ cannot have an effect on other factors. We know from interview investigations that an opinion (attitude) can be affected by the interviewer. The fact that a person is interviewed and answers ($M = 0$) may affect his attitude; if however a certain person should not be interviewed ($M = 1$) (perhaps, e.g., he cannot be contacted, is sick or is unwilling to answer) his attitude will not be affected, i.e. $M \not\rightarrow -y$. It would seem that such situations, where the act of interviewing itself affects the dependent variable, can be regarded as exceptions and they will not be treated here. For instance, in the example of the follow-up study in 1.1, we may regard the case where $M \not\rightarrow -y$ as being completely excluded.

### 3.8. Concluding remarks

The situations illustrated here are those for which we shall endeavor to find suitable solution approaches. We have described them as standard situations.[1]

The selection mechanisms in 3.61 are discussed in Chap. 4, but no new approaches to attack these problems are given. The mechanisms in 3.62 are treated in Chap. 4 and some new approaches are given in Chap. 5. Note that we have only attacked the situations where $x$ and $w$ are independent. The mechanism in 3.63 is treated in Chaps. 6 and 7.

[1] As for other problems, we have, for instance, given the example in Fig. 1 of section 2.32 where the problem is to distribute the effect of litter size on birth weight along two different "paths". In my terminology, this problem is not a selection problem.

CHAPTER 4

# INCOMPLETE DATA PROBLEMS

## WITH SPECIAL REFERENCE TO SITUATIONS WHERE DATA ON THE DEPENDENT VARIABLE ARE INCOMPLETE

### 4.0. Introduction

The aim of the present chapter is to give an account of certain incomplete data[1] problems—here denoted $M$-problems—and their statistical treatment. Some explanatory investigations with $M$-problems are introduced in section 4.1. A systematic discussion of the different mechanisms of the $M$-problems is given in section 4.2. A short survey of the literature concerning different methods of treating $M$-problems is set out in section 4.3. The new contributions follow in Chap. 5, which is built up on Chap. 4.

### 4.1. Some examples of $M$-problems

#### 4.11. *Samples from one group*

See the example given in (3.61 $\alpha$).

#### 4.12. *Samples from two groups (follow-up study)*

See the second example in (1.1).

#### 4.13. *Samples related to the regression problem (follow-up study)*

We wish to estimate the regression coefficient in the equation $y = \alpha + \beta x$, in a certain population, where $y$, for example, is the growth in grams from birth to 12 months of age, $x$ is the birthweight, and the population consists of pigs.

Assume that the birthweight is known for all the pigs, but that growth data up to 12 months of age are incomplete: the reason for the incomplete data being the death of some animals. Since it is more often the smallest animals who are slaughtered, e.g. the sick or the weak, it can easily be understood that an estimation of $\beta$ carried out on the material concerning the surviving pigs, would not generally correspond to an estimate based on material with completely known data.

---

[1] Including problems due to missing data or non-response.

### 4.14. *Samples from two groups*

Nyquist (1952) studied whether or not the different material used in the making of artificial dentures had an effect on the frequency of palatal ulcers. He asked everyone who had received complete dentures in the Royal school of dentistry in Stockholm during 1930–45 to participate in an investigation. Of these people (approx. 9000) about 15 % co-operated, and became the subjects for a further investigation, at which the denture material and the conditions of the palatal ulcers (denture sore mouth) were examined. The denture material was divided into two groups, rubber and acrylic, and in combination with the condition of the palatal ulcers it was then possible to make the following four-fold table.

TABLE

| Denture material | Without palatal ulcers | | With palatal ulcers | | Total | |
|---|---|---|---|---|---|---|
| | No. | % | No. | % | No. | % |
| Rubber | 305 | 71 | 125 | 29 | 430 | 100 |
| Acrylic | 537 | 82 | 123 | 18 | 660 | 100 |

Apparently the frequency of palatal ulcers with acrylic-dentures is 11 % lower than that of the other group.

In section 1.2 it has been noticed that a table of this kind will be difficult to interpret because of selective factors (such a factor is for example the age of the denture because all older dentures were made of rubber). We satisfy ourselves here by trying to find an answer to the following question: Of the population requested to co-operate in the investigation, how much higher or lower would the percentage of palatal ulcers be among those wearing acrylic-dentures than among those wearing rubber ones? In answering this question we have access only to a limited sample, which is certainly not random.

### 4.15. *Comments on the examples*

Common characteristics of the examples under sections 4.12–4.14:

(i) All the examples have incomplete data on material or measurement.

(ii) We wish to observe both an exaplanatory variable $x$ and a dependent variable $y$ within a certain population.

(iii) We wish to know something about the mean difference between two groups, $D$, about the regression coefficient $\beta_{yx}$ (or the correlation coefficient $\varrho_{yx}$ between two variables); in all cases parameters, referring to a certain population from which the present material has been obtained.

(iv) In all the examples we suspect that the non-response has given rise to a bias. If we calculate a sample estimate $D^*$, $\beta^*_{yx}$ or $\varrho^*_{yx}$ based on our observations, the mathematical expectation of this is not as a rule equal to that of $D$, $\beta_{yx}$ or $\varrho_{yx}$ respectively.

Bias on account of missing observations, is defined as the difference existing between the estimate obtained from observed material, and that which would have been obtained if information had been available for the whole population.

We also wish to point out an important difference in the degree of incompleteness between the examples. Here are two situations: (1) The data are incomplete in both the exaplanatory and the dependent variable (cf. ex. 4.14). (2) The data are incomplete only in the dependent variable; for the explanatory variable we have complete information about the whole population (cf. ex. 4.12 and 4.13).

## 4.2. Causal mechanism behind the missing observations

In Chap. 3 various selection problems have been discussed and different missing-data situations have been demonstrated by arrow schemes (3.6). In what follows we shall illustrate varying situations (a–d) which play a fundamental role both in the problem of assessing whether or not an $M$ bias is present, and also in that of building up various theoretical approaches to treat selection bias due to missing observations.

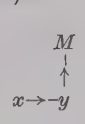
### 4.21. *M-situations with only x and y variables*

We begin with relatively simple situations, since only variables $x$, $y$ and $M$ have to be considered. When drawing up the arrow scheme we decide that it represents the original population, and that $M$ denotes the variable "missing – not missing" concerning the $y$ observation in the population. Arrows from $x$ and $y$ can be directed towards $M$, but never in the other direction (cf. 3.62).

The situations which are illustrated below can also be characterized by certain correlation coefficients calculated on the population having the mathematical expectation = 0, thereby assuming a linear relationship between the two variables. Various causal mechanisms:

*Situation* a: Some observations are missing, but neither $x$ nor $y$ have influenced $M$ in any way. This is *not* an $M$-situation which gives rise to bias. $\varrho_{yM.x} = 0$ and $\varrho_{yx.M} = 0$ holds good here ($\varrho$ refers to population).

*Situation* b. $M$ is affected by $y$ but not by $x$, i.e. $\varrho_{xM.y} = 0$. (In this situation $E(b_{yx}) \neq \beta_{yx}$ is valid as a rule, where $b_{yx}$ is an estimate based on the observed material.)




*Situation* c. $M$ is affected by $x$, but not by $y$, i.e. $\varrho_{yM.x} = 0$. This situation does not give rise to bias.

*Situation* d. $M$ is affected by both $x$ and $y$. $E(b_{yx}) \neq \beta_{yx}$; a bias is introduced.

Depending on the subject matter it is sometimes possible to allocate material to one or other of the situations a–d. Some examples are given below.

*Example of situation* a: $M$ can, for example, be caused by carelessness in chemical analysis (certain values become quite out of proportion and are discarded). $M$ often depends in such cases on a third factor which is completely independent of $x$ and $y$. This situation is common in experiments.

*Example of situation* b: In a chemical analysis low y values are discarded because such values appear to be out of proportion. If, on analysis of blood serum one finds Na-values higher than 125 mEq per litre, it is natural to suspect that mistakes have been made in the analysis, and therefore these values are discarded. The $x$ variable can, for example, be age, race or sex. It is plausible to conclude that the analytical error was caused by a factor which is independent of both $x$ and the true $Na$ value (Vahlquist-Melander *et al.*, 1959).

*Example of situation* c (cf. 4.14 and Nyquist, 1952): The palatal ulcer with which we are concerned in this investigation, is very difficult for the patient to discover since it is not likely to cause him any noticeable discomfort. From questions put to the patients taking part in the investigation it was made clear that they suffered the same personal difficulties from their dentures whether palatal ulcers were present or not. It is reasonable to conclude that the presence or absence of palatal ulcers ($y$) need not be considered to have any noticeable effect on $M$. The denture material, however, is probably associated with $M$; it is not unlikely that $\varrho_{xM} \neq 0$.

*Example of situation* d (cf. also example 4.13): We wish to estimate the differences in unemployment ($y$) between the sexes in a certain area between the years 1953–58, where $y$ = number of weeks employed, $x$ = sex.

The original population (i.e. people between the ages 25–60 in 1953) will decrease because, among other reasons some people leave the district, thus causing $M$. We may assume that one cause of this migration is the unemployment situation—one looks for a job elsewhere. It is also plausible to assume that men are more mobile than women. Therefore not only $x$, but also $y$ probably have an effect on $M$.

### 4.22. *M-situations where the w variable is present*

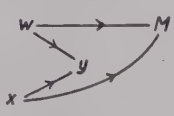
The situation is often further complicated by the presence of a completely hidden factor ($w$). Here we distinguish five situations:

Situation A: $w$ has an effect only on $x$, only on $y$ or only on $M$.
Situation B: $w$ has an effect on $x$ and $y$, but not on $M$, i.e. $\varrho_{wM.xy} = 0$.
Situation C: $w$ has an effect on $x$ and $M$, but not on $y$, i.e. $\varrho_{wy.x} = 0$.
Situation D: $w$ has an effect on $y$ and $M$, but not on $x$, i.e. $\varrho_{wx} = 0$.
Situation E: $w$ has an effect on $x$, $y$ and $M$.

These five situations A–E can be combined with a, b, c or d, and by so doing, altogether 20 combinations can exist. We shall try to treat A- and D-situations in chapter 5. We content ourselves here by pointing out that in situation A, $w$ can be interpreted as a residual factor, and that situation A d can therefore be written simply as d, having the same solution. As soon as we come to situation c D however we see that the presence of $w$ gives rise to a serious bias. The following example, introduced by Berkson illustrates this.

### 4.23. *Berkson's example on non-response via a hidden factor (variable)*

*Arrow scheme:*




$x = 0$: non-smoker; $x = 1$: smoker
$y = 0$: living; $y = 1$: dead
$w = 1$: Group I; $w = 2$: Group II.

Fig. 10.

According to this arrow scheme we see that: $\varrho_{xM.y} \neq 0$ as well as $\varrho_{wM.x} \neq 0$. Notice also that $\varrho_{xw} = 0$.

Berkson (1955) has illustrated how non-response could cause a spurious association. The example has been extensively commented upon in Wallis & Roberts (1956) which is quoted as follows:

> The population consists of two groups: Group I, constituting 3 percent of the population, which is on the verge of death and has a death rate during the period of the study of 99 percent; and Group II, constituting 97 percent of the population, which has a death rate of 0,03 percent. Assume that 80 percent of each group are smokers, and that the death rates are the same for smokers and nonsmokers within each group. Thus, the death rates would be (in percents):

| | Group I | Group II |
|---|---|---|
| Non-smokers | 99 | 0.03 |
| Smokers | 99 | 0.03 |

Notice that group I and group II cannot be directly observed. Data exist only for both groups collectively. The adjoining table accounts both for the original population, which cannot be observed, and also for a sample consisting of the response group.[1]

[1] What we here refer to as "sample" has been called remaining material in Chap. 5.

The table below is hypothetical and shows that the difference with respect to death-rate between smokers and non-smokers in the sample may be produced though no difference exists in the population.

*Hypothetical illustration of possible effects of selectivity in producing spurious association.*

(Observed data have been indicated here by an asterisk.)

| | Number in group | | | Number of deaths | | | |
|---|---|---|---|---|---|---|---|
| Smoker | Group I | Group II | Both groups | Group I | Group II | Both groups | Rate (%) |
| Population | | | | | | | |
| No | 600 | 19,400 | 20,000 | 594 | 6 | 600 | 3.0 |
| Yes | 2400 | 77,600 | 80,000 | 2376 | 24 | 2400 | 3.0 |
| Total | 3000 | 97,000 | 100,000 | 2970 | 30 | 3000 | 3.0 |
| Sample | | | | | | | |
| No | 300 | 19,206 | 19,506* | 297 | 6 | 303* | 1.55* |
| Yes | 1200 | 50,440 | 51,640* | 1188 | 16 | 1204* | 2.33* |
| Total | 1500 | 69,646 | 71,146* | 1485 | 22 | 1507* | 2.12* |

*Source:* Berkson, op. cit., Appendix Table 1.

The death rate of 3 % is the same for both the smokers and nonsmokers in the population. The sample on the other hand, shows that the smokers have a higher death-rate.

Berkson's example is characterized by a difference in non-response between smokers and non-smokers. The dependent variable however is not assumed to have any direct effect on the non-response. Indirectly the death-rate has an association with the non-response, since people with bad health (group I) are assumed to have both a high death-rate and a high nonresponse rate. In this connection the state of health is a hidden variable (cf. Wallis & Roberts, 1956, p. 288).

Berkson contents himself by pointing out that data about $x$ and $y$ in the non-response group are not sufficient to find out the genuine association between $x$ and $y$ in the population, i.e. $\varrho_{yx}$.

*Comments*

Although it is true that the present example is non-experimental, it is worth pointing out that the same difficulties would have arisen even if the investigation had been based on experimental material. If we imagine that in the above table the "no" group (non-smokers) corresponds to a control

group, and the "yes" group (smokers) to an experimental group, we then observe that the distribution of group I and group II is the same in the two population groups, smokers and non-smokers. Groups I and II play the same role as a controlled factor, neutralized by randomization at an experiment, in that $\varrho_{xw} = 0$ within the population.

### 4.24. *Supplementary observation data*

Starting from Berkson's illustration we can ask ourselves this question: Can the problem which arises because of $M$ be solved (i) by further previous knowledge about the non-response mechanism (compare above), (ii) by help of supplementary observation data as in the suggestions given below:

(*a*) With regard to the population we know the total number of living and also the total number of dead in both the "yes" and the "no" groups, i.e. 80,000 and 20,000 respectively. Although we know how many people were observed from the beginning these were not followed up. This situation often exists both in experimental and non-experimental follow-up investigations.

(*b*) We know the size of the population, 100,000, without however knowing the distribution between the "yes" and the "no" groups.

(*c*) We know the total number of both living and dead in the population, i.e. 97,000 and 3000 respectively without however knowing the proportion of smokers and non-smokers.

We should keep these suggestions in mind when attempting to build up new approaches to the problem.

## 4.3. A short survey of the literature

In 4.31–4.39 are presented some approaches to solution which can be found in the literature. Most of these have occurred in statistical application; the purely theoretical contributions seem to occur only rarely. This survey does not claim to be complete, and no attempt has been made to discover where the suggestions were first given.

It is difficult to give a survey of solutions which have been applied, because missing data problems are met in dealing with almost every kind of observation data and in almost every field of research.

It should be emphasized that this situation is not in accord with the ideal; i.e. having a *complete* observation data. From a practical point of view it can, however, be very expensive to complete the data and sometimes missing data seem to be unavoidable. After all, it cannot be denied that missing-data problems do exist, and that there is real need for methods suitable for treating these problems.

### 4.31. *Two-phase sampling*

The investigation is carried out in two phases. In the first phase, the aim is to obtain information concerning the whole sample. If this does not succeed one can try to complete missing data by a new—and often more intensive—collection of data from the missing-data group or a sample from this group (cf. Hurwitz and Hansen 1946, Cornfield 1942).

### 4.32. *Politz' and Simmons' method*

This method is very special and is only used in interview investigations (Politz 1949, Durbin 1954). For every interview, data are collected to estimate the risk of a callback. The method allows a rough classification of the non-responses.

### 4.33. *Comparison between those who have answered at once and those who have not answered until after a first or second request*

This method is used especially in mail investigations (see e.g. Hendricks, 1949). An application of the method is found in Moberg, 1951. Moberg compared income and success in university studies. He found that those who had first answered after the second request did not differ with respect to income from those who did not answer at all. These two groups differed considerably from those who had answered at once.

### 4.34. *Comparison between known attributes*

In some cases it is possible to compare attributes that are known in the response group as well as in the non-response group, e.g. age distribution, sex, income etc. (Cf. the comparison in the example in 4.33). If the difference with respect to these factors is small, it is indicated that the differences are small also with respect to the studied variables.

### 4.35. *The frequency distribution of the population is known*

Missing-data problems can sometimes be treated by the methods used for truncated distributions (cf. Hald, 1946).

### 4.36. *Special methods suitable in follow-up investigations*

An approach suitable in follow-up investigations where the original population is reduced, e.g. by deaths, has been discussed by Kimball (1957).

### 4.37. *Calculations of max. or min. values*

We assume that we know the size of the population. We know $\bar{y}$ for the remaining material and we know that $y$ can assume only certain finite values. We can then estimate a maximum and a minimum value for $\bar{y}_{pop}$ (=mean of $y$ for the whole population). This method, mentioned by Yule (1911), has been developed by Birnbaum-Sirken (1950) for the case where $y$ is assumed to be dichotomous. To use the method it is preferable that the $M$-group be small. In this case the difference between max. and min. values is relatively small. The method is illustrated in the following example.

We want to estimate the relative frequency of married people in a population. Our information may be summarized in the table below.

| | Married | Single | Total |
|---|---|---|---|
| Responses | 800 | 100 | 900 |
| Non-responses | ? | ? | 100 |

The relative frequency of married people within the populations can have the max. value $\frac{800+100}{1000}=90\,\%$ and the min. value $\frac{800}{1000}=80\,\%$.

The same approach is applicable even if we want to estimate the difference between two groups, e.g. men and women as in the table below.

| | Men | | Women | |
|---|---|---|---|---|
| | Married | Single | Married | Single |
| Responses | 440 | 60 | 350 | 50 |
| Non-responses | ? | ? | ? | ? |
| Total | 550 | | 450 | |

We denote married $y=1$ and single $y=0$. We now obtain for the population means:

$$\frac{440}{550}\leqslant\bar{y}_{men}<\frac{490}{550},\quad\frac{350}{450}\leqslant\bar{y}_{women}\leqslant\frac{400}{450}.$$

For the difference $d_{pop}$ between the frequencies within the population, we obtain

$$\frac{440}{550}-\frac{400}{450}\leqslant d_{pop}\leqslant\frac{490}{550}-\frac{350}{450}.$$

The max. difference here is positive and the min. difference negative. Under more favourable conditions we can get the same signs for both differences. The chance of getting the same signs, *ceteris paribus*, is greater, the less incomplete the data are.

CHAPTER 5

# NEW APPROACHES RELEVANT TO INCOMPLETE DATA

Approaches I, II and III are applicable if $y$ is a dichotomous variable. Approaches IV–V are relevant even if $y$ is continuous.

## 5.1. Estimation aspects of approaches I–III

*The aim of the estimation*

The following approaches intend to estimate:

(1) The difference, $D$, between two population groups, where $x = 1$ and $x = 0$ respectively, with regard to the mean value of the variable $y$:

$$D = \text{Population mean } (y|x = 1) - \text{Population mean } (y|x = 0).$$

(2) The ratio, $K$, between the population means for the same groups:

$$K = \frac{\text{Population mean } (y|x = 1)}{\text{Population mean } (y|x = 0)}.$$

(3) Weighted differences:

$$\bar{D} = \bar{D}_{v^{(i)}} = \frac{\sum v^{(i)} D^{(i)}}{\sum v^{(i)}},$$

where $(i)$ stands for a sub-population, $D^{(i)}$ is the difference within the sub-population $(i)$ between $y$-means of the groups, where $x = 1$ and $x = 0$ respectively, and $v^{(i)}$ is the weight attached to $D^{(i)}$.

The weights may, for instance, be chosen proportional to the number of individuals in the remaining material of the respective groups. The weighting system shall of course be chosen with due respect to the aim of the estimation. It often happens, however, that different weighting systems are tolerated, since the aim is somewhat vague.

(4) Weighted ratios:

$$\bar{K} = \bar{K}_{v^{(i)}} = \frac{\sum v^{(i)} K^{(i)}}{\sum v^{(i)}},$$

where $K^{(i)}$ is defined as the ratio referring to the $i$th sub-population and $v^{(i)}$ is the corresponding weight.

A weighting over different groups is often necessary if there exists a $w$-factor having an effect on $M$ as well as on $y$. In this situation we would prefer to estimate a difference or a ratio for every value of $w - w'$, $w''$, $w'''$

etc., but, due to the fact that the $w$-values are unknown, we must as a rule find a suitable weighting system and be content to characterize the difference between the two groups, $x = 1$ and $x = 0$ respectively, by only one coefficient.

## 5.2. Assumptions relevant to the approaches I–III. Survey

We try to estimate $D$, $\bar{D}_v$, $K$ or $\bar{K}_v$ when the material is incomplete as in the examples 4.12, 4.14 and 4.23. In these relatively simple cases we can find a consistent estimator, only if some supplementary or initial information exists. With the help of such information we can formulate assumption(s), varying from approach to approach. Later on we intend to discuss if and when these assumptions are realistic. We also present some applications (see pp. 134–139).

The following approaches are characterized by the supplementary information (assumptions) and the observed material:

The solutions summarized under the heading approach I (I$a$–I$l$) are based upon independent relations (see the arrow schemes) and observation data concerning $x$ and $y$.

Approach I$a$, I$d$ and I$e$: $x$ has no effect on $M$, only $x$ and $y$ are included.

Approach I$b$, I$f$ and I$g$: $y$ has no effect on $M$; only $x$ and $y$ are included.

Approach I$c$: $x$ and $y$ have no effect on $M$; a hidden variable $w$ independent of $x$ but having an effect on $M$ and $y$ exists and must be regarded.

Approach I$h$: $x$ has no effect on $M$; $y$ has an effect on $M$; $w$ has an effect on $y$ and $M$; $w$ is independent of $x$.

Approach I$k$: $x$ has an effect on $M$; $y$ has no effect on $M$; $w$ has an effect on $y$ and $M$; $w$ is independent of $x$.

Approach I$l$: A special approach, relevant if the frequency of missing data is the same in the group with $x = 0$ as in the group with $x = 1$.

Approach II: The estimation is carried out by means of an instrumental variable $z$. A hidden factor, $w$, is assumed to exist.

Approaches I and II are constructed upon certain independent relations.

Approach III: As a complement to the independent relations we assume that we have some special information concerning the mechanism causing the missing data.

## 5.3. Notation

The number of individuals (primary units) in the population have been denoted by $N$. For the corresponding numbers in the remaining material (sample) we use the symbol $n$. For sub-populations we use the superscripts

(′), (″), (*i*), e.g. N′, $N''$, $N^{(i)}$, and analogically in the remaining material. To denote different combinations of the values of $x$, $y$ and $z$, subscripts are used according to the system in the tables attached to approaches I*a* and II*a*, e.g. $N_{xy}$.

The inherent relation between these concepts: Number of subjects in the population = number of individuals in the remaining material + number of individuals in the missing-data material.

Number of individuals in the whole population = the sum of the numbers of individuals in the sub-populations.

The population mentioned above can even be regarded as a random sample from a larger population—a super-population.

If the $y$-values are disturbed by a random error they can also be regarded as randomly sampled from a super-population.

Both these cases can be treated as though the samples were identical with the population, as long as only large samples are considered.

The arrow-scheme presentation always corresponds to the aggregated case (cf. p. 28).

*Remark:* In all approaches we assume that the random effect, caused by the fact that the population or the sample is small, can be ignored, so that our estimators will be consistent.

We assume in approaches I, II and III that $x$ and $y$ are dichotomous variables. We let "$M = 1$" mean that $y$-data are missing, and "$M = 0$" mean that $y$-data are complete.

## 5.4. Approach I

In the following approaches we refer to the assumptions below:

(a) All individuals are observed with respect to $x$.

(b) Only the remaining individuals are observed with respect to $y$.[1]

(c) $y$ has (may have) an effect on $M$.

(d) $y$ has no effect on $M$.

(e) $x$ has (may have) an effect on $M$.

(f) $x$ has *no* other effect on $M$ than via $y$.

(g) A factor, $w$, having an effect on $x$, $y$ and $M$ exists.

(h) No factor having an effect on both $x$ and $y$ or on both $x$ and $M$ or on both $y$ and $M$ exists (or must be considered).

(j) A factor, $w$, having an effect on $y$ and $M$ exists.

[1] $y$-data are given for the individuals in the remaining the material but not for the individuals in the missing data material.

The assumptions (d), (f), (h) imply independence. No approach where (g) is assumed to hold is discussed. The factor $w$ is assumed to be underlying in relation to $y$ and/or $M$, and independent of $x$. $w$ which is never observed stands for a certain factor or a complex of factors.

The assumptions are illustrated by arrow schemes, which are always valid for the population.

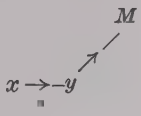
### *Approach I a*

*Assumptions:* (a), (b), (c), (f) and (h).

According to these assumptions the causal mechanism can be illustrated as follows.

*Arrow scheme*

(Cf. 4.21, situation b.):




*Material:* The table below illustrates how the population is distributed with respect to the four possible combinations of the factors $x$ and $y$.

| Combination of factors $x$ $y$ | Number of individuals in the population | Observed material $M=0$ | |
|---|---|---|---|
| 0 0 | $N_0$ | $T_0$ | $n_0$ |
| 0 1 | $N_y$ | | $n_y$ |
| 1 0 | $N_x$ | $T_x$ | $n_x$ |
| 1 1 | $N_{xy}$ | | $n_{xy}$ |

*Aim: Difference estimation.*

We want to estimate the difference, $D$, in the population given. $D$ is the difference between the relative frequency of $y$ for $x=1$ and the corresponding relative frequency for $x=0$, i.e.

$$D = N_{xy}/(N_x + N_{xy}) - N_y/(N_0 + N_y).$$

We can estimate $D$ if every one of the numbers $N_{xy}$, $N_y$, $N_x$ and $N_0$ is known. According to assumption (a) the following sums are observed:

$$N_0 + N_y = T_0; \tag{1}$$

$$N_x + N_{xy} = T_x. \tag{2}$$

By means of the assumption (f), we can derive the following expressions valid for $y=0$ and $y=1$ respectively:

$$\frac{n_0}{N_0} = \frac{n_x}{N_x} \quad \text{or} \quad \frac{n_0}{n_x} = \frac{N_0 - n_0}{N_x - n_x} = \frac{N_0}{N_x}; \tag{3}$$

$$\frac{n_y}{N_y} = \frac{n_{xy}}{N_{xy}} \quad \text{or} \quad \frac{n_y}{n_{xy}} = \frac{N_y - n_y}{N_{xy} - n_{xy}} = \frac{N_y}{N_{xy}}. \tag{4}$$

This means that four equations exist for estimating the unknown $N$'s in the missing-data material.

Every $n$-value—$n_0$, $n_y$, $n_x$, $n_{xy}$—can be observed directly; we know all combinations between $x$ and $y$ in the remaining material, i.e. when $M = 0$.

By means of these observations and equations (1)–(4) we can estimate $N_0$, $N_x$, $N_y$ and $N_{xy}$, i.e. all unknown numbers in the population. We obtain:

$$N_0 = \frac{n_{xy} T_0 - n_y T_x}{n_{xy} n_0 - n_y n_x} n_0; \tag{5}$$

$$N_x = \frac{n_{xy} T_0 - n_y T_x}{n_{xy} n_0 - n_y n_x} n_x; \tag{6}$$

$$N_y = \frac{n_0 T_x - n_x T_0}{n_{xy} n_0 - n_y n_x} n_y; \tag{7}$$

$$N_{xy} = \frac{n_0 T_x - n_x T_0}{n_{xy} n_0 - n_y n_x} n_{xy}. \tag{8}$$

We can now express $D$ in known quantities. After some simplifications we get:

$$D = \left( \frac{n_{xy}}{T_x} - \frac{n_y}{T_0} \right) \frac{n_0 T_x - n_x T_0}{n_{xy} n_0 - n_y n_x}. \tag{9}$$

*Remark:* A solution is obtained according to this formula if $n_{xy} n_0 - n_y n_x \neq 0$, i.e. if $x$ has an effect on $y$. If $x$ has no effect on $y$, $n_{xy} n_0 - n_y n_x = 0$.

*Aim: Ratio estimation.*

If we had intended to estimate

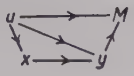
$$K = \frac{N_{xy}/(N_x + N_{xy})}{N_y/(N_0 + N_y)} \tag{10}$$

instead of $D$, we should find by means of eqs. (5)–(8) that

$$K = \frac{N_{xy}}{N_y} \frac{T_0}{T_x} = \frac{n_{xy}}{n_y} \frac{T_0}{T_x}.$$

This estimator, of course, applies to a problem different from that of estimating $D$.

*Remark:* In addition to $x$ we often have to consider other observed variables also; e.g. in the arrow scheme below, $u$ is a variable observed both if $M = 1$ and $M = 0$.




Let us consider a fixed value $u = u'$. Given this fixed value, the arrow scheme can be reduced to the same scheme as the one mentioned above.

We obtain an estimator for the difference, $D_{u'}$, corresponding to $u'$. In the same way, we can for $u = u''$ estimate a difference $D_{u''}$, etc. (cf. eq. (9)).

### *Approach I b*

*Assumptions:* (b), (d), (e) and (h).

Compared with app. I*a*, app. I*b* differs in that respect that $x$ has an effect on $M$ but $y$ has no effect on $M$.

*Arrow scheme*

(cf. 4.21 situation c):

*Solution:* From the assumptions it follows:

$$n_0/N_0 = n_y/N_y, \tag{11}$$

because $y$ has no effect on $M$ if $x = 0$.

$$n_x/N_x = n_{xy}/N_{xy}, \tag{12}$$

because $y$ has no effect on $M$ if $x = 1$.

The following numbers are known from the material: $n_0$, $n_x$, $n_y$, $n_{xy}$.

*Aim: Difference estimation.*

By means of these numbers and eq. (11) and eq. (12), $D$ can be estimated. If $D$ is expressed in known quantities we obtain

$$D = \frac{n_{xy}}{n_x + n_{xy}} - \frac{n_y}{n_0 + n_y}.$$

Observe that the missing data do not bias the estimation. If we had calculated a difference based only on the remaining material, the same estimator would have been obtained.

*Aim: Ratio estimation.*

For $K$ we find by substitution in eq. (11) and eq. (12) that

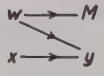
$$K = \frac{N_{xy}/(N_x + N_{xy})}{N_y/(N_0 + N_y)} = \frac{n_{xy}/(n_x + n_{xy})}{n_y/(n_0 + n_y)}.$$

## *Approach I c*

*Assumptions:* (b), (d), (f) and (j).

We start from a model, which in addition to $x$ and $y$ contains a hidden variable $w$ having an effect on $y$ and $M$. Moreover $w$ and $x$ are assumed to be independent. Further we assume that neither $x$ nor $y$ has an effect on $M$.

*Arrow scheme:*




*Notation:* Different values of $w$: $w'$, $w''$, $w^{(i)}$ etc. By $n_0^{(i)}$ is denoted that group in the remaining material for which $x = 0$, $y = 0$ and $w = w^{(i)}$, i.e. $\sum n_0^{(i)} = n_0$. Corresponding notations are applied to $n_x$, $n_y$, $n_{xy}$, $N_0$, $N_x$, $N_y$ and $N_{xy}$. By $D^{(i)}$ is meant the following difference in the sub-population $(i)$ for which $w$ assumes a fixed value $w = w^{(i)}$:

$$D^{(i)} = \frac{N_{xy}^{(i)}}{N_x^{(i)} + N_{xy}^{(i)}} - \frac{N_y^{(i)}}{N_0^{(i)} + N_y^{(i)}}. \tag{13}$$

*Solution:* For a fixed value of $w = w'$, it is assumed that neither $x$ nor $y$ has an effect on $M$. It follows that

$$\frac{n_0'}{N_0'} = \frac{n_x'}{N_x'} = \frac{n_y'}{N_y'} = \frac{n_{xy}'}{N_{xy}'}. \tag{14'}$$

In general for $w = w^{(i)}$, it follows that

$$\frac{n_0^{(i)}}{N_0^{(i)}} = \frac{n_x^{(i)}}{N_x^{(i)}} = \frac{n_y^{(i)}}{N_y^{(i)}} = \frac{n_{xy}^{(i)}}{N_{xy}^{(i)}}. \tag{14$^{(i)}$}$$

Since $x$ and $w$ are independent:

$$\frac{N_0' + N_y'}{N_x' + N_{xy}'} = \frac{N_0'' + N_y''}{N_x'' + N_{xy}''} = \frac{N_0^{(i)} + N_y^{(i)}}{N_x^{(i)} + N_{xy}^{(i)}}. \tag{15}$$

According to eq. (14′) it follows:

$$\frac{n_0' + n_y'}{n_x' + n_{xy}'} = \frac{N_0' + N_y'}{N_x' + N_{xy}'}, \tag{16′}$$

and analogically we obtain:

$$\frac{n_0^{(i)} + n_y^{(i)}}{n_x^{(i)} + n_{xy}^{(i)}} = \frac{N_0^{(i)} + N_y^{(i)}}{N_x^{(i)} + N_{xy}^{(i)}}. \tag{16$^{(i)}$}$$

We introduce the symbol $c$ and deduce from eq. (15) and eq. (16). Since $\sum n_0^{(i)} = n_0$, $\sum n_y^{(i)} = n_y$ etc. we find:

$$\frac{n_0' + n_y'}{n_x' + n_{xy}'} = \frac{n_0^{(i)} + n_y^{(i)}}{n_x^{(i)} + n_{xy}^{(i)}} = \frac{n_0 + n_y}{n_x + n_{xy}} = c.$$

*Aim: Difference estimation.*

From eq. (14) we find that $D^{(i)} = n_{xy}^{(i)}/(n_x^{(i)} + n_{xy}^{(i)}) - n_y^{(i)}/(n_0^{(i)} + n_y^{(i)})$.

We construct the difference, $\bar{D}$, by weighting the differences $D'$, $D''$, $D^{(i)}$ etc. with respect to the frequency of *remaining* individuals for different values of $w$ (we introduce corresponding weights, denoted $n_{w^{(i)}}$):

$$\bar{D} = \frac{\sum n_{w^{(i)}} D^{(i)}}{\sum n_{w^{(i)}}} = \frac{\sum n_{w^{(i)}} \left( \frac{N_{xy}^{(i)}}{N_x^{(i)} + N_{xy}^{(i)}} - \frac{N_y^{(i)}}{N_0^{(i)} + N_y^{(i)}} \right)}{\sum n_{w^{(i)}}};$$

i.e.

$$\bar{D} = \frac{\sum n_{w^{(i)}} \left( \frac{n_{xy}^{(i)}}{n_x^{(i)} + n_{xy}^{(i)}} - \frac{n_y^{(i)}}{n_y^{(i)} + n_0^{(i)}} \right)}{\sum n_{w^{(i)}}}, \tag{17}$$

where $n_{w^{(i)}}$ is defined as

$$n_{w^{(i)}} = n_0^{(i)} + n_x^{(i)} + n_y^{(i)} + n_{xy}^{(i)}. \tag{18}$$

We find from eqs. (16) and (18):

$$n_{w^{(i)}} = (c+1)\,(n_x^{(i)} + n_{xy}^{(i)}) \tag{19}$$

Substituting $n_{w^{(i)}}$ according to eq. (19) we find that eq. (17) is reduced to:

$$\bar{D} = \frac{\sum \left( n_{xy}^{(i)} - \frac{n_y^{(i)}}{c} \right)(c+1)}{\sum n_{w^{(i)}}} = \frac{\frac{c+1}{c}\,(c\,n_{xy} - n_y)}{n_0 + n_x + n_y + n_{xy}}.$$

Since $c = (n_0 + n_y)/(n_x + n_{xy})$, we deduce the estimator

$$\bar{D} = \frac{n_{xy}}{n_x + n_{xy}} - \frac{n_y}{n_0 + n_y}.$$

$\bar{D}$ can evidently be estimated without considering the missing data, by using only the remaining material.

*Remark*: If $D' = D'' = D^{(i)}$ etc. we find that

$$D = \frac{n_{xy}}{n_x + n_{xy}} - \frac{n_y}{n_0 + n_y}$$

is independent of the weighting system; $n_{w^{(i)}}$ can here be chosen arbitrarily.

*Aim: Ratio estimation.*

We construct $$\bar{K}_{v^{(i)}} = \frac{\sum v^{(i)} K^{(i)}}{\sum v^{(i)}},$$

where $K^{(i)}$, corresponding to $w = w^{(i)}$, is defined as

$$K^{(i)} = \frac{N_{xy}^{(i)} / (N_x^{(i)} + N_{xy}^{(i)})}{N_y^{(i)} / (N_0^{(i)} + N_y^{(i)})}. \tag{20}$$

After some simplification, we find:

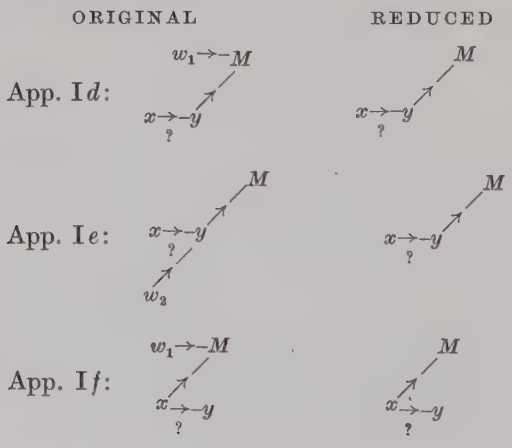
$$K^{(i)} = \frac{N_{xy}^{(i)}}{N_y^{(i)}} \frac{n_0^{(i)} + n_y^{(i)}}{n_x^{(i)} + n_{xy}^{(i)}} = \frac{N_{xy}^{(i)}}{N_y^{(i)}} \frac{n_0 + n_y}{n_x + n_{xy}}.$$

If we now choose $v^{(i)} = N_y^{(i)}$, we obtain

$$\bar{K}_{v^{(i)}} = \bar{K}_{N_y^{(i)}} = \frac{\sum N_{xy}^{(i)}}{\sum N_y^{(i)}} \frac{n_0 + n_y}{n_x + n_{xy}} = \frac{N_{xy}}{N_y} \frac{n_0 + n_y}{n_x + n_{xy}} = \frac{n_{xy}}{n_y} \frac{n_0 + n_y}{n_x + n_{xy}}.$$

*Approaches I d–I g.*

*Arrow schemes:*

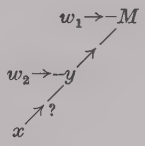


App. I*g*:


*Solutions of approaches I d and I e:*

In app. I*d* and app. I*e* the *assumptions* are the same as those in app. I*a*. However, we have added the factors $w_1$ and $w_2$ respectively, which are assumed to be independent of $x$. $D$ and $K$ can be estimated in the same way as in app. I*a*.

These estimators are even valid if we combine the arrow schemes I*d* and I*e* as follows:

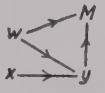



We here assume that $w_1$ and $w_2$ are independent.

*Solutions of approaches I f and I g:*

In app. I*f* and app. I*g* the *assumptions* are the same as those in app. I*b*. The factors $w_1$ and $w_2$ are assumed to be independent of $x$. The same solution as that used in app. I*b* is valid.

### *Approach I h.*

*Arrow scheme:* 


*Assumptions:* (a), (b), (c), (f) and (j).
*Notation:* See app. I*c*.
*Solution:* Since $x$ has no effect on $M$,

$$\frac{n_0'}{N_0'} = \frac{n_x'}{N_x'}, \quad \frac{n_y'}{N_y'} = \frac{n_{xy}'}{N_{xy}'}, \quad \frac{n_0^{(i)}}{N_0^{(i)}} = \frac{n_x^{(i)}}{N_x^{(i)}} \quad \text{etc.} \tag{21}$$

Further, $w$ and $x$ being independent,

$$\frac{N_0' + N_y'}{N_x' + N_{xy}'} = \frac{N_0'' + N_y''}{N_x'' + N_{xy}''} = \frac{N_0^{(i)} + N_y^{(i)}}{N_x^{(i)} + N_{xy}^{(i)}} = \frac{T_0}{T_x}. \tag{22}$$

In eq. (22) $T_0 = \sum (N_0^{(i)} + N_y^{(i)}) = N_0 + N_y$ and $T_x = \sum (N_x^{(i)} + N_{xy}^{(i)}) = N_x + N_{xy}$. Note that $T_0$ and $T_x$ are known (cf. assumption a).

*Ratio estimation.*

We construct the ratio, $K^{(i)}$, defined in eq. (20). According to eq. (22) and eq. (21) we find:

$$K^{(i)} = \frac{N_{xy}^{(i)}}{N_y^{(i)}} \frac{T_0}{T_x} = \frac{n_{xy}^{(i)}}{n_y^{(i)}} \frac{T_0}{T_x} \quad \text{etc.} \tag{23}$$

We now choose as weights $n_y'$, $n_y''$ and $n_y^{(i)}$ respectively. We obtain as an estimator of $\bar{K}$ the following expression:

$$\bar{K} = \frac{\sum n_y^{(i)} \frac{N_{xy}^{(i)}}{N_y^{(i)}} \frac{T_0}{T_x}}{\sum n_y^{(i)}} = \frac{n_{xy}}{n_y} \frac{T_0}{T_x}. \tag{24}$$

Thus, the same estimator would have been obtained if the missing data had been ignored.

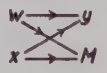
The weighting system is, however, somewhat arbitrary. If we had been forced to work with another weighting system, it would had been difficult to find a solution.

*Remark:* If the weights $n_y^{(i)}$ were substituted by $N_y^{(i)}$ the same estimator, $\bar{K}$, would have been obtained.

### *Approach I k.*

*Assumptions:* (a), (b), (d), (e), (j) and according to eq. (30).

*Arrow scheme:*




Note that $w$ has no effect on $x$.

*Notation* (see app. I c and I a): $q_0^{(i)}$, $q_x^{(i)}$ and $\Delta$ are defined in eq. (25) and (27).

*Solution:* From the assumptions it follows, because $y$ has no effect on $M$:

$$\frac{n_0^{(i)}}{N_0^{(i)}} = \frac{n_y^{(i)}}{N_y^{(i)}} = q_0^{(i)}, \quad \frac{n_x^{(i)}}{N_x^{(i)}} = \frac{n_{xy}^{(i)}}{N_{xy}^{(i)}} = q_x^{(i)}. \tag{25}$$

Since $x$ and $w$ are independent, we obtain:

$$\frac{N_x' + N_{xy}'}{N_0' + N_y'} = \frac{N_x'' + N_{xy}''}{N_0'' + N_y''} = \frac{N_x^{(i)} + N_{xy}^{(i)}}{N_0^{(i)} + N_y^{(i)}} = \frac{T_x}{T_0}. \tag{26}$$

*Aim: Estimation of a weighted difference.*

The difference, $D^{(i)}$ for the population group for which $w = w^{(i)}$ is defined as in eq. (13).

A weighted difference, called $\Delta$, is now defined:

$$\Delta = \frac{\sum (n_0^{(i)} + n_x^{(i)} + n_y^{(i)} + n_{xy}^{(i)})\, D^{(i)}}{\sum (n_0^{(i)} + n_x^{(i)} + n_y^{(i)} + n_{xy}^{(i)})}, \tag{27}$$

i.e. the weights are proportional to the number of remaining observations in respective population groups; the same weights as those in app. I*c* have been chosen. Our aim is to estimate $\Delta$. We find after some substitutions:

$$\Delta = \frac{n_{xy} - n_y + \sum \left( N_{xy}^{(i)}\, q_0^{(i)} \frac{T_0}{T_x} - N_y^{(i)}\, q_x^{(i)} \frac{T_x}{T_0} \right)}{n_0 + n_x + n_y + n_{xy}}$$

or

$$\Delta = \frac{n_0 - n_x + \sum \left( N_0^{(i)}\, q_x^{(i)} \frac{T_x}{T_0} - N_x^{(i)}\, q_0^{(i)} \frac{T_0}{T_x} \right)}{n_0 + n_x + n_y + n_{xy}}.$$

Further the following relations are valid:

$$\sum N_{xy}^{i}\, q_x^{i} = n_{xy}; \quad \text{every } N^{(i)} \geqslant 0; \tag{28}$$

$$\sum N_y^{i}\, q_0^{i} = n_y; \quad \text{every } q^{(i)} \geqslant 0 \text{ and } q^{(i)} \leqslant 1. \tag{29}$$

*Further assumptions:* Let us further assume that the expression

$$(q_0^{(i)} - q_x^{(i)}) \tag{30}$$

has the same sign for every $(i)$.

We substitute the relation (30) in the expressions for $\Delta$. It is then possible to find a maximum and minimum estimate for $\Delta$.

Maximum values are:

$$\frac{(T_x + T_0)(n_{xy} T_0 - n_y T_x)}{T_x T_0 (n_0 + n_x + n_y + n_{xy})} \quad \text{if} \quad q_x^{(i)} \geqslant q_0^{(i)},$$

$$\frac{(T_x + T_0)(n_0 T_x - n_x T_0)}{T_x T_0 (n_0 + n_x + n_y + n_{xy})} \quad \text{if} \quad q_x^{(i)} \leqslant q_0^{(i)}.$$

Minimum values are:

$$\frac{(T_x + T_0)(n_0 T_x - n_x T_0)}{T_x T_0 (n_0 + n_x + n_y + n_{xy})} \quad \text{if} \quad q_x^{(i)} \geqslant q_0^{(i)},$$

$$\frac{(T_x+T_0)(n_{xy}T_0-n_yT_x)}{T_xT_0(n_0+n_x+n_y+n_{xy})} \quad \text{if} \quad q_x^{(i)} \leqslant q_0^{(i)}.$$

*Remark:*

$$\text{If} \quad \frac{n_x+n_{xy}}{T_x}-\frac{n_0+n_y}{T_0} \quad \begin{cases} >0, & \text{then} \quad q_x^{(i)} \geqslant q_0^{(i)}, \\ <0, & \text{then} \quad q_x^{(i)} \leqslant q_0^{(i)}. \end{cases} \tag{31}$$

In other words, if the difference, $q_x^{(i)}-q_0^{(i)}$ has the same sign for every $(i)$, we can by means of the above inequalities conclude that $q_x^{(i)} \geqslant q_0^{(i)}$ or $q_x^{(i)} \leqslant q_0^{(i)}$.

### *Approach I l.*

This approach is useful if we observe that the missing data is as frequent for the group $x=0$ as for the group $x=1$. We intend to study how to use this information if the same arrow scheme as that in app. I $a$ or in app. I $h$ is relevant.

From the material we have found that

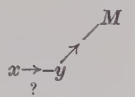
$$\frac{n_0+n_y}{N_0+N_y}=\frac{n_x+n_{xy}}{N_x+N_{xy}}. \tag{32}$$

We construct the same $D^{(i)}$ as that in eq. (13).

For the remaining material we define:

$$d^*=\frac{n_{xy}}{n_x+n_{xy}}-\frac{n_y}{n_0+n_y} \tag{33}$$

*Arrow scheme*
(cf. app. I $a$):




If (32) is observed then $y \rightarrow - M$ can exist only if $\bar{D}=0$; in that case $d^*=0$. If $\bar{D} \neq 0$ the arrow between $y$ and $M$ may be eliminated, it follows that $\bar{D}=d^*$, and therefore that $d^*$ is an estimator.

*Arrow scheme*
(cf. app. I $h$):

Since $x$ has no effect on $M$, we have (cf. eq. (21)):

$$\left.\begin{aligned} &\frac{n_0'}{N_0'}=\frac{n_x'}{N_x'}=q_0', \qquad \frac{n_y'}{N_y'}=\frac{n_{xy}'}{N_{xy}'}=q_y'; \\ &\frac{n_0^{(i)}}{N_0^{(i)}}=\frac{n_x^{(i)}}{N_x^{(i)}}=q_0^{(i)}, \quad \frac{n_y^{(i)}}{N_y^{(i)}}=\frac{n_{xy}^{(i)}}{N_{xy}^{(i)}}=q_y^{(i)}. \end{aligned}\right\} \tag{34}$$

Since $x$ and $w$ are independent we find (cf. eq. (22)):

$$(N_0^{(i)} + N_y^{(i)})\,(N_x^{(j)} + N_{xy}^{(j)}) = (N_0^{(j)} + N_y^{(j)})\,(N_x^{(i)} + N_{xy}^{(i)}) \quad \text{for} \quad i \neq j. \tag{35}$$

We *assume* further:

$$\text{sign}\ (q_y^{(i)} - q_0^{(i)}) = \text{sign}\ (q_y^{(j)} - q_0^{(j)}). \tag{36}$$

$$\text{sign}\ D^{(j)} = \text{sign}\ D^{(i)}. \tag{37}$$

By substituting $q_0^{(i)}$ and $q_y^{(i)}$ we obtain from eq. (32):

$$\frac{\sum N_0^{(i)}\, q_0^{(i)} + N_y^{(i)}\, q_y^{(i)}}{N_0 + N_y} = \frac{\sum N_x^{(i)}\, q_0^{(i)} + N_{xy}^{(i)}\, q_y^{(i)}}{N_x + N_{xy}},$$

$$\text{i.e.} \quad \sum q_y^{(i)} \left( \frac{N_{xy}^{(i)}}{N_x + N_{xy}} - \frac{N_y^{(i)}}{N_0 + N_y} \right) - \sum \left( \frac{N_0^{(i)}}{N_0 + N_y} - \frac{N_x^{(i)}}{N_x + N_{xy}} \right) q_0^{(i)} = 0. \tag{38}$$

Since

$$D^{(i)} = \frac{N_{xy}^{(i)}}{N_x^{(i)} + N_{xy}^{(i)}} - \frac{N_y^{(i)}}{N_0^{(i)} + N_y^{(i)}} = \frac{N_0^{(i)}}{N_0^{(i)} + N_y^{(i)}} - \frac{N_x^{(i)}}{N_x^{(i)} + N_{xy}^{(i)}},$$

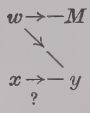
eq. (38) can be written:

$$\sum q_y^{(i)}\, D^{(i)} - \sum q_0^{(i)}\, D^{(i)} = \sum (q_y^{(i)} - q_0^{(i)})\, D^{(i)} = 0.$$

If eq. (36) as well as eq. (37) are to be valid, then necessarily either $q_y^{(i)} = q_0^{(i)}$ or $D^{(i)} = 0$.

We can thus divide the material into two groups:

I. The group of individuals for which $q_y^{(h)} = q_0^{(h)}$ has the arrow scheme:




II. The group of individuals for which $D^{(j)} = 0$, has the arrow scheme:

Let us now construct $\bar{D}$ as in app. I*c*. Eq. (13) is still valid. Eq. (16) will also be valid, for if $D^{(j)} = 0$, then necessarily

$$\frac{N_0^{(j)} + N_y^{(j)}}{N_x^{(j)} + N_{xy}^{(j)}} = \frac{n_0^{(j)}}{n_x^{(j)}} = \frac{n_y^{(j)}}{n_{xy}^{(j)}} = \frac{n_0^{(j)} + n_y^{(j)}}{n_x^{(j)} + n_{xy}^{(j)}} \tag{39}$$

and if $q_0^{(h)} = q_y^{(h)}$, then necessarily

$$\frac{N_0^{(h)} + N_y^{(h)}}{N_x^{(h)} + N_{xy}^{(h)}} = \frac{n_0^{(h)} + n_y^{(h)}}{n_x^{(h)} + n_{xy}^{(h)}}. \tag{40}$$

From this it follows that eq. (16) is true. Furthermore, eqs. (15), (17) and (19) must be true.

From this it follows that

$$\bar{D} = \frac{n_{xy}}{n_x + n_{xy}} - \frac{n_y}{n_0 + n_y}.$$

i.e. $d^*$ (cf. eq. 33) is an estimator of $\bar{D}$.

## 5.5. Approach II

These approaches illustrate how a solution is reached by means of an instrumental variable, $z$, assuming that $z$ has no effect on $M$.

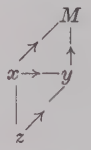
### Approach II *a*

*Assumptions and observed material:*

(a) data exist for all individuals concerning the variables $x$ and $z$ even in the missing material (see table below);
(b) corresponding data for $y$ exist only in the remaining material;
(c) $x$ may have an effect on $M$ and $y$;
(d) $z$ has no other effect on $M$ than via $y$ or $x$;
(e) $z$ has an effect on $y$;
(f) no other factor(s) exists having an effect on more than one of the factors illustrated in the arrow scheme;
(g) $y$ is a dichotomous variable.

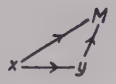
As no factors are assumed to have an effect on $x$, the present approach would be relevant in an experiment where $x$ is varied by the experimenter.

*Arrow scheme:* From the assumptions valid for the population this arrow scheme follows:




Below all possible factor combinations—including $x$, $z$, $y$ and $M$—are presented. The same notation as that in app. I is used. The relation between $x$ and $z$ may be unilateral, either $x$ or $z$ being an explaining factor, or bilateral.

*Remark:* It is not possible to obtain a general solution in the situation below, unless one is given further information or assumptions.




If we have a variable $z$, corresponding to the assumptions above, it is possible to find a solution.

TABLE

| Combination of factors $x$ $z$ $y$ $M$ | Sub-population groups | Observation material | |
|---|---|---|---|
| 0 0 0 0 | $n_0$ | $n_0$ | $T_0$ |
| 0 0 1 0 | $n_y$ | $n_y$ | |
| 0 0 0 1 | $N_0 - n_0$ | | |
| 0 0 1 1 | $N_y - n_y$ | | |
| 1 0 0 0 | $n_x$ | $n_x$ | $T_x$ |
| 1 0 1 0 | $n_{xy}$ | $n_{xy}$ | |
| 1 0 0 1 | $N_x - n_x$ | | |
| 1 0 1 1 | $N_{xy} - n_{xy}$ | | |
| 0 1 0 0 | $n_z$ | $n_z$ | $T_z$ |
| 0 1 1 0 | $n_{zy}$ | $n_{zy}$ | |
| 0 1 0 1 | $N_z - n_z$ | | |
| 0 1 1 1 | $N_{zy} - n_{zy}$ | | |
| 1 1 0 0 | $n_{xz}$ | $n_{xz}$ | $T_{xz}$ |
| 1 1 1 0 | $n_{xzy}$ | $n_{xzy}$ | |
| 1 1 0 1 | $N_{xz} - n_{xz}$ | | |
| 1 1 1 1 | $N_{xzy} - n_{xzy}$ | | |

For the remaining material, i.e. for $M = 0$, data are known for some factor combinations, namely $n_0$, $n_y$, $n_x$, $n_{xy}$, $n_z$, $n_{zy}$, $n_{xz}$ and $n_{xzy}$. According to assumption (a) we have data for $T_0$, $T_x$, $T_z$ and $T_{xz}$. Observe that the notations $N_0$, $N_x$, $N_y$, $N_{xy}$, $n_0$, $n_x$, $n_y$, $n_{xy}$ in the table above, have been reserved for factor combinations where $z = 0$.

*Aim and solution:* We shall try to reconstruct the number of individuals in all the sixteen sub-population groups in the table. We must consider only eight groups, since we already have full information concerning eight groups belonging to the remaining material. Furthermore, four equations are clear from the table, namely:

$$N_0 + N_y = T_0, \tag{41}$$

$$N_x + N_{xy} = T_x, \tag{42}$$

$$N_z + N_{zy} = T_z, \tag{43}$$

$$N_{xz} + N_{xzy} = T_{xz}. \tag{44}$$

To these can be added four additional equations, since according to assumption (d), $z$ has no direct effect on $M$:

$$\frac{n_0}{n_z} = \frac{N_0}{N_z} \quad (x=0, y=0), \tag{45}$$

$$\frac{n_x}{n_{xz}} = \frac{N_x}{N_{xz}} \quad (x=1, y=0), \tag{46}$$

$$\frac{n_y}{n_{zy}} = \frac{N_y}{N_{zy}} \quad (x=0, y=1), \tag{47}$$

$$\frac{n_{xy}}{n_{xzy}} = \frac{N_{xy}}{N_{xzy}} \quad (x=1, y=1). \tag{48}$$

From these eight equations it is possible to estimate the unknown $N$-values. The following solutions are obtained:

$$\left.\begin{aligned}
N_y &= \frac{n_0 T_z - n_z T_0}{n_0 n_{zy} - n_y n_z} n_y, & N_0 &= T_0 - N_y; \\
N_{xy} &= \frac{n_x T_{xz} - n_{xz} T_x}{n_x n_{xzy} - n_{xy} n_{xz}} n_{xy}, & N_x &= T_x - N_{xy}; \\
N_{zy} &= \frac{n_0 T_z - n_z T_0}{n_0 n_{zy} - n_y n_z} n_{zy}, & N_z &= T_z - N_{zy}; \\
N_{xzy} &= \frac{n_x T_{xz} - n_{xz} T_x}{n_x n_{xzy} - n_{xy} n_{xz}} n_{xzy}, & N_{xz} &= T_{xz} - N_{xzy}.
\end{aligned}\right\} \tag{49}$$

*Remark:* Eqs. (41)–(48) give a solution if

$$n_0 n_{zy} - n_y n_z \neq 0, \quad \text{i.e. if } N_0 N_{zy} - N_y N_z \neq 0,$$

and if

$$n_x n_{xzy} - n_{xy} n_{xz} \neq 0, \quad \text{i.e. if } N_x N_{xzy} - N_{xy} N_{xz} \neq 0.$$

These inequalities are valid according to assumption (e).

The effect of $x$ on $y$ can be estimated for $z = 0$ and for $z = 1$. For $z = 0$ we define the difference $D(z=0)$ as

$$D\,(z=0) = N_{xy}/(N_{xy}+N_x) - N_y/(N_y+N_0).$$

Analogically we define $D\,(z=1)$ as

$$D\,(z=1) = N_{xzy}/(N_{xzy}+N_{xz}) - N_{zy}/(N_{zy}+N_z).$$

We obtain after substitutions:

$$D\,(z=0) = \frac{n_x\,T_{xz} - n_{xz}\,T_x}{n_x\,n_{xzy} - n_{xy}\,n_{xz}}\,\frac{n_{xy}}{T_x} - \frac{n_0\,T_z - n_z\,T_0}{n_0\,n_{zy} - n_y\,n_z}\,\frac{n_y}{T_0}.$$

For $z=1$ we estimate analogically $D\,(z=1)$:

$$D\,(z=1) = \frac{n_x\,T_{xz} - n_{xz}\,T_x}{n_x\,n_{xzy} - n_{xy}\,n_{xz}}\,\frac{n_{xzy}}{T_{xz}} - \frac{n_0\,T_z - n_z\,T_0}{n_0\,n_{zy} - n_y\,n_z}\,\frac{n_{zy}}{T_z}.$$

Then $D(z=0)$ and $D(z=1)$ can be weighted together to give a single estimator.

*Can this solution method be generalized?*

If $y$ had more than two values, would it be possible to find a solution as was done in app. II*a*? Let $y$ have $V_y$ different values, and $x$ have $V_x$ different values and $z$ have $V_z$ different values. The arrow scheme above is still assumed to be valid. Counting the number of equations, we find that the difference between the number of unknown equations and the number of known ones, is:

$$\underset{\text{(unknown)}}{V_y\,V_x\,V_z} - \underset{\text{(cf. eq. 41–44)}}{V_x\,V_z} - \underset{\text{(cf. eq. 45–48)}}{V_x\,V_y\,(V_z-1)} = V_x\,(V_y - V_z).$$

A solution is possible only if $z$ assumes as many or more discrete values than $y$.

**Approach II *b***

*Assumptions:* If we assume that $y$ has an effect on $z$ rather than the contrary, app. II*a* is still applicable.

*Arrow scheme:*

In this case we gain nothing by estimating $D(z=0)$ or $D(z=1)$ separately. A more reasonable aim would be to estimate:

$$D = \frac{N_{xy}+N_{xzy}}{N_x+N_{xz}+N_{xy}+N_{xzy}} - \frac{N_y+N_{zy}}{N_0+N_z+N_y+N_{zy}}.$$

The estimators for the eight $N$-values given above are still valid. We obtain:

$$D = \frac{(n_x T_{xz} - n_{xz} T_x)(n_{xy} + n_{xzy})}{(T_x + T_{xz})(n_x n_{xzy} - n_{xy} n_{xz})} - \frac{(n_0 T_z - n_z T_0)(n_y + n_{zy})}{(T_0 + T_z)(n_0 n_{zy} - n_y n_z)}. \qquad (50)$$

**Approaches II *c* and II *d***

*Assumptions:* Like those in app. II *a*, except that the factors $w_1$ and $w_2$ are added.

*Arrow scheme:*

App. II *c*:


App. II *d*:


App. II *a* is applicable in II *c* as well as in II *d*. The same is true if we superimpose II *c* on II *d*, assuming that $w_1$ and $w_2$ are independent.

**Approach II *e***

We assume as in app. I *l* that the relative frequency of missing data is the same for $z = 0$ and $z = 1$.

Moreover, we assume that $z$ has the same position as $x$ in app. I *h* and that the following *arrow scheme* is relevant:

*Assumptions and notations:*

(a) The relative frequency of missing data is the same for $z = 0$ and $z = 1$:

$$\frac{n_0 + n_y}{N_0 + N_y} = \frac{n_z + n_{zy}}{N_z + N_{zy}} \text{ and } \frac{n_x + n_{xy}}{N_x + N_{xy}} = \frac{n_{xz} + n_{xzy}}{N_{xz} + N_{xzy}} \text{ are valid.}$$

(b) By $d_0'$, $d_0''$, and $d_0^{(i)}$ we denote the direct *z-effect on y* in sub-populations, where $x = 0$ and $w$ assumes the values $w'$, $w''$ and $w^{(i)}$ respectively. The notations $d_x'$, $d_x''$ and $d_x^{(i)}$ are analogous, corresponding to sub-populations where $x = 1$. We define $d_0^{(i)}$ and $d_x^{(i)}$:

$$d_0^{(i)} = \frac{N_{zy}^{(i)}}{N_{zy}^{(i)} + N_z^{(i)}} - \frac{N_y^{(i)}}{N_y^{(i)} + N_0^{(i)}}; \qquad d_x^{(i)} = \frac{N_{xzy}^{(i)}}{N_{xzy}^{(i)} + N_{xz}^{(i)}} - \frac{N_{xy}^{(i)}}{N_{xy}^{(i)} + N_x^{(i)}}.$$

We assume that $d_0'$, $d_0''$, $d_0^{(i)}$ etc. have the *same sign* and are *not equal to* 0 and that $d_x'$, $d_x''$, $d_x^{(i)}$ have the same sign and are not equal to 0.

(c) The following *notations* are introduced:

$$q_y^{(i)} = \frac{n_y^{(i)}}{N_y^{(i)}} = \frac{n_{zy}^{(i)}}{N_{zy}^{(i)}}; \quad q_0^{(i)} = \frac{n_0^{(i)}}{N_0^{(i)}} = \frac{n_z^{(i)}}{N_z^{(i)}};$$

$$q_{xy}^{(i)} = \frac{n_{xy}^{(i)}}{N_{xy}^{(i)}} = \frac{n_{xzy}^{(i)}}{N_{xzy}^{(i)}}; \quad q_x^{(i)} = \frac{n_x^{(i)}}{N_x^{(i)}} = \frac{n_{xz}^{(i)}}{N_{xz}^{(i)}}.$$

We assume that $q_y' - q_0'$, $q_y'' - q_0''$, $q_y^{(i)} - q_0^{(i)}$ etc. have the same sign; and that $q_{xy}' - q_x'$, $q_{xy}'' - q_x''$, $q_{xy}^{(i)} - q_x^{(i)}$ also have the same sign.

*Proof:*

(Analogous to the proof in I *l*.)

Let us first consider the case where $x = 0$. The factor $x$ being fixed, the arrow scheme is reduced to:

From the given assumptions it follows then that $q_y^{(i)} = q_0^{(i)}$ (cf. app. I *l*); note that $d_0^{(j)} \neq 0$ according to the assumption (b). It is then evident that the arrow connecting $y$ and $M$ can be eliminated. In the same way—$x = 1$ being fixed—it can be proved that $q_x^{(i)} = q_{xy}^{(i)}$. Hence the arrow scheme can be simplified; note that even $z$ can be eliminated:

This system is in accordance with app. I *k*. The correctness of the simplification of the arrow scheme has been possible to prove by introducing an instrumental variable, $z$. This method, of course, is usable only if the arrow, $y \rightarrow -M$, does not exist in reality. Moreover, it is necessary to find a suitable instrumental variable.

*Remark:* The same arguments are applicable also if $z$ has an effect on $x$.

## 5.6. Approach III

We have not hitherto, except in app. II *e*, attempted to attack the incomplete-data problems in the following situation:

We intend to prove that these problems can also be attacked by introducing a new variable, $z$. We must, however, add some new assumptions which are more complicated than those used in app. II$a$. This makes it difficult to find out if the assumptions are realistic or not.

The new assumptions are of the type discussed in section 2.65.

In the following approaches we start from the arrow scheme:

*Observation material:* The $y$-values are known only for the remaining material. The $w$-values are totally unknown. The $x$-values and $z$-values are completely known for the whole population.

## Approach III$a$

*Notations and assumptions:*

Notations of population subgroups.

| | | $z=0$ | | | | $z=1$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| $x$ | $y$ | $w=w'$ | $w=w''$ | $w=w^{(i)}$ | Total | $w=w'$ | $w=w''$ | $w=w^{(i)}$ | Total |
| 0 | 0 | $N_0'$ | $N_0''$ | $N_0^{(i)}$ | $N_0$ | $N_z'$ | $N_z''$ | $N_z^{(i)}$ | $N_z$ |
| 0 | 1 | $N_y'$ | $N_y''$ | $N_y^{(i)}$ | $N_y$ | $N_{zy}'$ | $N_{zy}''$ | $N_{zy}^{(i)}$ | $N_{zy}$ |
| 1 | 0 | $N_x'$ | $N_x''$ | $N_x^{(i)}$ | $N_x$ | $N_{xz}'$ | $N_{xz}''$ | $N_{xz}^{(i)}$ | $N_{xz}$ |
| 1 | 1 | $N_{xy}'$ | $N_{xy}''$ | $N_{xy}^{(i)}$ | $N_{xy}$ | $N_{xzy}'$ | $N_{xzy}''$ | $N_{xzy}^{(i)}$ | $N_{xzy}$ |

$N$ specified with a subscript and a superscript refers to a sub-population and $n$ with corresponding indices refers to the remaining material in the corresponding sub-population.

According to the arrow scheme we derive the following relations. New notations $q'$, $q''$, $q^{(i)}$ etc. for the remaining data frequency are introduced:

$$\left.\begin{aligned}
&\frac{n_0'}{N_0'} = \frac{n_z'}{N_z'} = q_0', \quad \frac{n_0^{(i)}}{N_0^{(i)}} = \frac{n_z^{(i)}}{N_z^{(i)}} = q_0^{(i)};\\
&\frac{n_y'}{N_y'} = \frac{n_{zy}'}{N_{zy}'} = q_y', \quad \frac{n_y^{(i)}}{N_y^{(i)}} = \frac{n_{zy}^{(i)}}{N_{zy}^{(i)}} = q_y^{(i)};\\
&\frac{n_x'}{N_x'} = \frac{n_{xz}'}{N_{xz}'} = q_x', \quad \frac{n_x^{(i)}}{N_x^{(i)}} = \frac{n_{xz}^{(i)}}{N_{xz}^{(i)}} = q_x^{(i)};\\
&\frac{n_{xy}'}{N_{xy}'} = \frac{n_{xzy}'}{N_{xzy}'} = q_{xy}', \quad \frac{n_{xy}^{(i)}}{N_{xy}^{(i)}} = \frac{n_{xzy}^{(i)}}{N_{xzy}^{(i)}} = q_{xy}^{(i)}.
\end{aligned}\right\} \qquad (51)$$

Since $x$ and $w$ are independent in the population:

$$\frac{N_0' + N_y'}{N_x' + N_{xy}'} = \frac{N_0'' + N_y''}{N_x'' + N_{xy}''} = \frac{N_0^{(i)} + N_y^{(i)}}{N_x^{(i)} + N_{xy}^{(i)}} = \frac{T_0}{T_x}. \tag{52}$$

Analogically, since $z$ and $w$ are independent, we obtain:

$$\frac{N_0' + N_y'}{N_z' + N_{zy}'} = \frac{N_0'' + N_y''}{N_z'' + N_{zy}''} = \frac{N_0^{(i)} + N_y^{(i)}}{N_z^{(i)} + N_{zy}^{(i)}} = \frac{T_0}{T_z}, \tag{53}$$

and

$$\frac{N_0' + N_y'}{N_{xz}' + N_{xzy}'} = \frac{N_0'' + N_y''}{N_{xz}'' + N_{xzy}''} = \frac{N_0^{(i)} + N_y^{(i)}}{N_{xz}^{(i)} + N_{xzy}^{(i)}} = \frac{T_0}{T_{xz}}. \tag{54}$$

We add the following *new assumptions* ($q_y$, $q_x$, $q_{xy}$ are new notations):

$$\frac{q_y'}{q_0'} = \frac{q_y''}{q_0''} = \frac{q_y^{(i)}}{q_0^{(i)}} = q_y, \tag{55}$$

$$\frac{q_x'}{q_0'} = \frac{q_x''}{q_0''} = \frac{q_x^{(i)}}{q_0^{(i)}} = q_x, \tag{56}$$

$$\frac{q_{xy}'}{q_0'} = \frac{q_{xy}''}{q_0''} = \frac{q_{xy}^{(i)}}{q_0^{(i)}} = q_{xy}. \tag{57}$$

These assumptions would be reasonable, if each of the variables $x$, $y$, and $w$ were—independently of the others—capable of giving rise to missing (or remaining) data (cf. 2.65).

*Estimation of a weighted difference.*

We now obtain:

$$D' = \frac{N_{xy}'}{N_x' + N_{xy}'} - \frac{N_y'}{N_0' + N_y'} = \frac{1}{N_0' + N_y'} \left( \frac{N_{xy}' \, T_0}{T_x} - N_y' \right)$$

$$= \frac{1}{q_{xy}' \, (N_0' + N_y') \, T_x} \left( n_{xy}' \, T_0 - n_y' \, T_x \frac{q_{xy}}{q_y} \right),$$

$$D^{(i)} = \frac{1}{q_{xy}^{(i)} \, (N_0^{(i)} + N_y^{(i)}) \, T_x} \left( n_{xy}^{(i)} \, T_0 - n_y^{(i)} \, T_x \frac{q_{xy}}{q_y} \right).$$

From eqs. (51), (55) and (57) we obtain $q_{xy}^{(i)} \, (N_y^{(i)} + N_0^{(i)}) = \frac{q_{xy}}{q_y} n_y^{(i)} + q_{xy} \, n_0^{(i)}$. This expression can be substituted in the denominator of $D^{(i)}$.

Moreover we obtain from eqs. (51) and (52) that

$$\frac{n_0'/q_0' + n_y'/q_y'}{n_x'/q_x' + n_{xy}'/q_{xy}'} = \frac{n_0^{(i)}/q_0^{(i)} + n_y^{(i)}/q_y^{(i)}}{n_x^{(i)}/q_x^{(i)} + n_{xy}^{(i)}/q_{xy}^{(i)}} = \frac{T_0}{T_x}.$$

This expression can be simplified by means of (55), (56) and (57). We obtain the following expression:

$$\frac{q_y n_0' + n_y'}{(q_y/q_x) n_x' + (q_y/q_{xy}) n_{xy}'} = \frac{q_y n_0^{(i)} + n_y^{(i)}}{(q_y/q_x) n_x^{(i)} + (q_y/q_{xy}) n_{xy}^{(i)}} = \frac{T_0}{T_x}.$$

From this it follows by summations of the numerators and denominators:

$$\frac{q_y n_0 + n_y}{(q_y/q_x) n_x + (q_y/q_{xy}) n_{xy}} = \frac{T_0}{T_x}. \tag{58}$$

In the same way we obtain from eqs. (51), (53), (54), (55) and (57) that

$$\frac{q_y n_0 + n_y}{q_y n_z + n_{zy}} = \frac{T_z}{T_0}, \tag{59}$$

and from (51), (54), (55), (56) and (57) that

$$\frac{q_y n_0 + n_y}{(q_y/q_x) n_{xz} + (q_y/q_{xy}) n_{xzy}} = \frac{T_0}{T_{xz}}. \tag{60}$$

From these three expressions we can estimate $q_y$, $q_{xy}$ and $q_x$.
As weights for $D'$, $D''$, and $D^{(i)}$ we choose: $v'$, $v''$ and $v^{(i)}$ respectively as below, where $v' + v'' + \cdots + v^{(i)} + \cdots + \cdots = 1$. We define

$$\frac{v'}{(q_{xy}/q_y) n_y' + q_{xy} n_0'} = \frac{v''}{(q_{xy}/q_y) n_y'' + q_{xy} n_0''} = \frac{v^{(i)}}{(q_{xy}/q_y) n_y^{(i)} + q_{xy} n_0^{(i)}} = \lambda.$$

These weights are chosen proportionally to the denominators of $D^{(i)}$.
From this expression it follows:

$$\lambda = \frac{1}{(q_{xy}/q_y) n_y + q_{xy} n_0}.$$

We estimate: $D^* = v' D' + v'' D'' + v''' D''' + \cdots + v^{(i)} D^{(i)} + \cdots +$

$$= \frac{n_{xy} T_0 - n_y T_x (q_{xy}/q_y)}{T_x ((q_{xy}/q_y) n_y + q_{xy} n_0)} = \frac{n_{xy} T_0 q_y - n_y T_x q_{xy}}{T_x (n_y + q_y n_0)} \frac{1}{q_{xy}} \tag{61}$$

As $q_{xy}$, $q_x$, and $q_y$ can be obtained from eqs. (58), (59) and (60) it is possible to calculate $D^*$.

We obtain:

$$q_y = \frac{T_0 n_{zy} - T_z n_y}{T_z n_0 - T_0 n_z}$$

$$q_x = \frac{n_x n_{xzy} - n_{xy} n_{zx}}{n_{zy} n_0 - n_z n_y} \cdot \frac{T_z n_0 - T_0 n_z}{T_x n_{xzy} - T_{xz} n_{xy}} \cdot \frac{T_0 n_{zy} - T_z n_y}{T_z n_0 - T_0 n_z}$$

$$q_{xy} = \frac{n_x n_{xzy} - n_{xy} n_{xz}}{n_{zy} n_0 - n_z n_y} \cdot \frac{T_0 n_{zy} - T_z n_y}{T_{xz} n_x - T_x n_{xz}}.$$

*General comments on approach III a.*

(i) The variable $z$ is necessary in order to obtain a solution.

(ii) In every system where $w$ has an effect on $M$ as well as on $y$, the solution is valid for only special weighting systems.

(iii) The solution given above is relevant only to the population groups for which $z = 0$. A solution for $z = 1$ can be constructed analogically.

**Approach III b**

*Assumptions and notations.*

The same notations as in app. III $a$.

In relation to app. III $a$, the assumptions eqs. (55), (56) and (57) are substituted by:

$$\frac{N_0'}{N_y'} = \frac{N_0''}{N_y''} = \frac{N_0^{(i)}}{N_y^{(i)}} = \frac{N_0}{N_y},$$

$$\frac{N_x'}{N_{xy}'} = \frac{N_x''}{N_{xy}''} = \frac{N_x^{(i)}}{N_{xy}^{(i)}} = \frac{N_x}{N_{xy}},$$

$$\frac{N_z'}{N_{zy}'} = \frac{N_z''}{N_{zy}''} = \frac{N_z^{(i)}}{N_{zy}^{(i)}} = \frac{N_z}{N_{zy}},$$

$$\frac{N_{xz}'}{N_{xzy}'} = \frac{N_{xz}''}{N_{xzy}''} = \frac{N_{xz}^{(i)}}{N_{xzy}^{(i)}} = \frac{N_{xz}}{N_{xzy}}.$$

From eqs. (51)–(54) and the above assumptions we obtain:

$$\frac{n_0'}{n_z'} = \frac{n_0''}{n_z''} = \frac{n_0^{(i)}}{n_z^{(i)}} = \frac{n_0}{n_z} = \frac{N_0}{N_z},$$

$$\frac{n_x'}{n_{xz}'} = \frac{n_x''}{n_{xz}''} = \frac{n_x^{(i)}}{n_{xz}^{(i)}} = \frac{n_x}{n_{xz}} = \frac{N_x}{N_{xz}},$$

$$\frac{n_y'}{n_{zy}'} = \frac{n_y''}{n_{zy}''} = \frac{n_y^{(i)}}{n_{zy}^{(i)}} = \frac{n_y}{n_{zy}} = \frac{N_y}{N_{zy}},$$

$$\frac{n_{xy}'}{n_{xzy}'} = \frac{n_{xy}''}{n_{xzy}''} = \frac{n_{xy}^{(i)}}{n_{xzy}^{(i)}} = \frac{n_{xy}}{n_{xzy}} = \frac{N_{xy}}{N_{xzy}}.$$

The last mentioned equalities correspond to eqs. (45), (46), (47), and (48) respectively in app. II*a*. Hence a solution analogous to app. II*a* can be obtained which holds for every weighting system since $D^{(i)} = D^{(j)}$.

**Approach III *c***

*Assumptions and arrow scheme:* If $y$ has no effect on $M$, i.e.

a solution can be obtained if the same assumptions as those in app. III*a* are valid, but it is not necessary to use any $z$-variable (i.e. assumptions (51), (53), and (54) are superfluous). The same notations as in III*a* are used.

As $y$ has no effect on $M$ we can formulate the new assumptions:

$$\frac{n_0'}{N_0'} = q_0' = \frac{n_y'}{N_y'} = q_y'; \quad \frac{n_0^{(i)}}{N_0^{(i)}} = q_0^{(i)} = \frac{n_y^{(i)}}{N_y^{(i)}} = q_y^{(i)} \tag{62}$$

and

$$q_x' = q_{xy}'; \quad q_x^{(i)} = q_{xy}^{(i)} \quad \text{etc.}$$

The expressions $q_y$, $q_{xy}$, and $q_x$ given in (55), (56), and (57) are simplified in accordance with eq. (62): $q_y = 1$ and $q_{xy} = q_x$.

By means of eqs. (51) and (62), eq. (52) is reduced to

$$\frac{n_0 + n_y}{(1/q_x)\, n_x + (1/q_{xy})\, n_{xy}} = \frac{T_0}{T_x}, \quad \text{i.e.} \quad q_{xy} = q_x = \frac{T_0}{T_x} \frac{n_x + n_{xy}}{n_0 + n_y}.$$

From this expression $q_x$ can be estimated.

*Estimation of a weighted difference.*

We construct $D'$, $D''$, $D^{(i)}$ etc. as was done in app. III*a*, and simplify by means of eq. (62):

$$D^{(i)} = \frac{N_{xy}^{(i)}}{N_x^{(i)} + N_{xy}^{(i)}} - \frac{N_y^{(i)}}{N_0^{(i)} + N_y^{(i)}} = \frac{n_{xy}^{(i)}}{n_x^{(i)} + n_{xy}^{(i)}} - \frac{n_y^{(i)}}{n_0^{(i)} + n_y^{(i)}}.$$

We choose the following weights (cf. app. I*c*)

$$n_{w'} = n_0' + n_x' + n_y' + n_{xy}',$$
$$n_{w^{(i)}} = n_0^{(i)} + n_x^{(i)} + n_y^{(i)} + n_{xy}^{(i)} \quad \text{etc.},$$

and attempt to estimate:

$$\Delta = \frac{n_{w'} D' + n_{w''} D'' + n_{w^{(i)}} D^{(i)} + \cdots}{n_{w'} + n_{w''} + n_{w^{(i)}} + \cdots}$$

$$= \frac{\sum \left\{ \frac{q_0^{(i)}}{q_{xy}^{(i)}} n_{xy}^{(i)} \frac{N_0 + N_y}{N_x + N_{xy}} \right\} - n_y + n_{xy} - \sum n_y^{(i)} \frac{q_x^{(i)}}{q_y^{(i)}} \frac{N_x + N_{xy}}{N_0 + N_y}}{n_0 + n_y + n_x + n_{xy}}$$

$$= \frac{\frac{1}{q_x} n_{xy} \frac{N_0 + N_y}{N_x + N_{xy}} - n_y + n_{xy} - q_x n_y \frac{N_x + N_{xy}}{N_0 + N_y}}{n_0 + n_y + n_x + n_{xy}}.$$

We obtain after substituting $q_x$:

$$\Delta = \frac{n_{xy}}{n_x + n_{xy}} - \frac{n_y}{n_0 + n_y}.$$

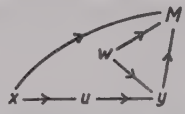
If we had ignored the missing-data problem and naively based our estimate on the remaining material we would have obtained the same result as above. Observe that the weighting system is the same as that in app. I*c*.

The arrow scheme is like that in app. I*k*, but the assumptions are more special than those in app. I*k*. On the other hand, app. III*c* gives a point estimator of $\Delta$. In app. I*k* we can only find min. or max. values for $\Delta$.

**Approach III *d***

In this approach we introduce an *intervening* variable, $u$.

*Arrow scheme:*




*Assumptions:* All assumptions can be read from the arrow scheme.

*Observation material:* For all subjects we have data concerning $x$, $u$, and $M$, but data concerning $y$ for only the remaining material.

*Aim: Estimation of* $\bar{K}$.

We want to estimate the same $\bar{K}$ as on p. 70.

*Solution method:* According to the arrow scheme $x$ has no effect on $y$ other than via $u$. We can therefore study the effect of $x$ in two steps: (i) the effect of $x$ on $y$, (ii) the effect of $u$ on $y$. We can limit ourselves to estimating the effect of $u$ on $y$. This estimator can be obtained as was done in app. I*h*.

### 5.7–5.8. Approaches IV–V

Two $x$-groups, $x = 0$ and $x = 1$, shall be compared, but in contrast to approaches I–III we assume that the $y$-values are not dichotomous or restricted to limited values.

*Observations:* The material is complete with regard to $x$ but incomplete with regard to $y$.

*Notations:* The two $x$-groups are denoted by the subscripts $(x=)\,0$ and $(x=)\,1$ respectively. For the $y$-values we introduce the symbol $\overline{Y}$ for population means, and $Y^{\text{Med}}$ for population medians. The relative frequency of remaining material where $x=0$ is denoted $R_{x=0}(y)$ and is a function of $y$. By $R_{x=1}(y)$ is denoted the corresponding relative frequency where $x=1$. Of course the following relations must be true:

$$0 \leqslant R_{x=0}(y) \leqslant 1 \text{ and } 0 \leqslant R_{x=1}(y) \leqslant 1.$$

By $\phi_{x=1}(y)$ and $\phi_{x=0}(y)$ we denote the cumulative relative frequency distribution for the whole population, given $x=1$ and $x=0$ respectively.

By $p_{x=1}$ we mean the relative frequency of individuals in the remaining material belonging to the population group where $x=1$, and by $p_{x=0}$ the corresponding relative frequency of individuals in the remaining material belonging to the population group where $x=0$. By the ratio $F_{x=1}(y)/p_{x=1}$ we mean the cumulative relative frequency distribution with respect to $y$ in the remaining material where $x=1$, and by $F_{x=0}(y)/p_{x=0}$ we define the corresponding relative frequency distribution where $x=0$.

*The aim of the estimation:*

1) To estimate the difference between population means with respect to $y$ for $x=1$ and $x=0$ respectively, i.e.:

$$D = \overline{Y}_{x=1} - \overline{Y}_{x=0}.$$

2) To estimate the difference $\Delta$ between population medians, for $x=1$ and $x=0$, with respect to $y$:

$$\Delta = Y^{\text{Med}}_{x=1} - Y^{\text{Med}}_{x=0}.$$

*A general solution.*

It is now possible to study the difference $\Delta = Y^{\text{Med}}_{x=1} - Y^{\text{Med}}_{x=0}$ in a rather simple way by finding a min. and a max. value.

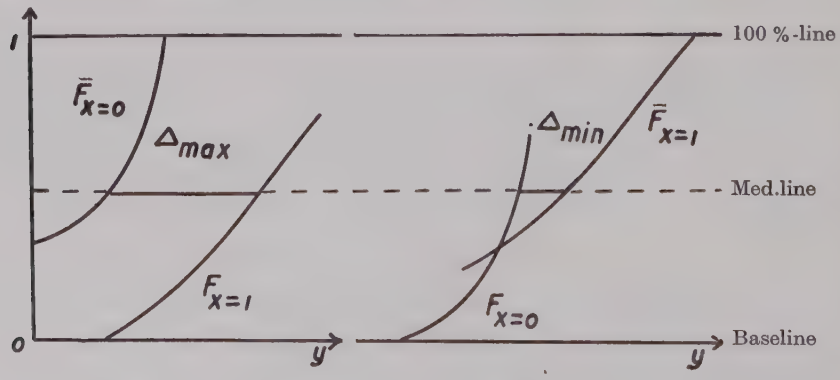



Fig. 11.




Fig. 12.

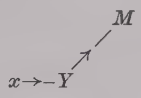
$\Delta_{max}$ and $\Delta_{min}$ correspond to the whole-drawn part of the median line.

$F_{x=1}$ in Fig. 11 (and $F_{x=0}$ in Fig. 12) starts from the base line and reaches a frequency amounting to $p_{x=1}$ (and $p_{x=0}$ respectively). By $\overline{F}_{x=0}$ and $\overline{F}_{x=1}$ we mean: $\overline{F}_{x=0}(y) = F_{x=0}(y) + 1 - p_{x=0}$ and $\overline{F}_{x=1}(y) = F_{x=1}(y) + 1 - p_{x=1}$. $\overline{F}_{x=0}$ and $\overline{F}_{x=1}$ are thus parallel to $F_{x=0}$ and $F_{x=1}$ respectively. Both $\overline{F}_{x=0}$ and $\overline{F}_{x=1}$ reach the 100 %-line in the figures.

From the figures we can see that $p_{x=1} > \frac{1}{2}$ and $p_{x=0} > \frac{1}{2}$. We cannot find a greater value for $\Delta$ than that represented in Fig. 11; the remaining material cannot be placed in a more extreme position. The same is true for Fig. 12; we cannot replace the material so that a lesser value of $\Delta$ can be obtained. These solutions hold for the difference between medians; but we cannot find a limited value for the mean difference, $D$, if $y$ itself is not assumed to be limited in some way. These arguments point out an important fact: a median solution is often possible where a finite mean solution does not exist. By the given example it can even be shown that $\Delta_{\max}$ and $\Delta_{\min}$ can both be positive, but this, of course, is not a general rule.

## 5.7. Approach IV

*Arrow scheme:*




Our basic assumption is that $x$ has no effect on $M$; i.e. that $R_{x=0}(y) = R_{x=1}(y) = R(y)$.

To this assumption we add one or more assumptions from the following list:

(a) $R(y)$ is a non-increasing function of $y$.

(b) $R(y)$ is a non-decreasing function of $y$.

(c) $\phi_{x=0}(y) \geqslant \phi_{x=1}(y)$ for every $y$.

(d) $\phi_{x=0}(y) \leqslant \phi_{x=1}(y)$ for every $y$.

In the present section we intend to demonstrate by some case studies how combinations of these assumptions can be used to obtain a solution.

App. IV*a*: We assume that (a) is valid and that $F_{x=0}(y) \geqslant F_{x=1}(y)$.
IV*b*: We assume that (b) is valid and that $\overline{F}_{x=0}(y) \geqslant \overline{F}_{x=1}(y)$.
IV*c*: We assume that either (a) or (b) is valid.
IV*d*: We assume that (c) is valid.
IV*e*: We assume that (d) is valid.
IV*f*: We assume that either (c) or (d) is valid.
IV*g*: We assume that (a) and (c) are valid.

**Approach IV *a***

*Estimation of a max. value for the median difference.*

In proving that a max. value solution exists we shall limit ourselves to a special case. We add the assumption that from the observation material we know that

$F_{x=0}(y) \geqslant F_{x=1}(y)$ for all $y$ corresponding to $F_{x=0}(y) < \frac{1}{2}$

and that $p_{x=1} > \frac{1}{2}$ and $p_{x=0} > \frac{1}{2}$.

Let us choose the $y$-scale so that the lowest value in the observation material is $y = 0$; i.e. $F_{x=0}(y < 0) = 0$. We now construct

$$F_{x=1}(k\Delta) - F_{x=1}((k-1)\Delta) = f_{x=1}(k\Delta) = f_1(k\Delta)$$

and

$$F_{x=0}(k\Delta) - F_{x=0}((k-1)\Delta) = f_{x=0}(k\Delta) = f_0(k\Delta)$$

where $\Delta$ is positive and small, $x$ can be $= 0$ or $= 1$ and $f_x$, $f_1$ and $f_0$ are frequency distributions. The difference

$$\sum_{k=0}^{n} f_{x=1}(k\Delta) - \sum_{k=0}^{n} f_{x=0}(k\Delta) < 0, \text{ when the corresponding difference}$$

$$F_{x=1}(n\Delta) - F_{x=0}(n\Delta) < 0.$$

Let us denote the frequency function corresponding to $\phi_x$ by $\varphi_x$.
If $R > 0$ we obtain that:

$$\frac{f_x(k\Delta)}{R(k\Delta)} = \varphi_x(k\Delta).$$

We know from the assumption (a) that $R$ is non-increasing, i.e. we know that

$$R((k+1)\Delta) - R(k\Delta) \leqslant 0.$$

Hence let us write

$$\frac{1}{R(k\Delta)} = r + \sum_{i=1}^{k} \theta_i,$$

where $\theta_i \geqslant 0$.

We introduce $n_0$ and $n_x$ so that the following relations hold:

$$\sum_{k=0}^{n_0} f_0(k\Delta) = \sum_{k=0}^{n_x} f_1(k\Delta) \leqslant \tfrac{1}{2}.$$

Here it must be the case that $n_0 < n_x$, since $\sum_{k=0}^{n} f_1(k\Delta) - \sum_{k=0}^{n} f_0(k\Delta) < 0$.

We obtain by substitution:

$$\phi_{x=0}(n_0\Delta) = \sum_{k=0}^{n_0} \frac{f_0(k\Delta)}{R(k\Delta)} = \sum_{k=0}^{n_0} \left\{ f_0(k\Delta) \left[ r + \sum_{i=1}^{k} \theta_i \right] \right\}$$

$$= \sum_{k=0}^{n_0} f_0(k\Delta)\, r + \theta_1 \sum_{k=1}^{n_0} f_0(k\Delta) + \theta_2 \sum_{k=2}^{n_0} f_0(k\Delta) + \cdots + \cdots + \theta_i \sum_{k=i}^{n_0} f_0(k\Delta) +$$

$$+ \cdots + \cdots + \theta_{n_0} f_0(n_0\Delta) \quad \text{and}$$

$$\phi_{x=1}(n_x\Delta) = \sum_{k=0}^{n_x} \frac{f_1(k\Delta)}{R(k\Delta)} = \sum_{k=0}^{n_x} f_1(k\Delta)\, r + \theta_1 \sum_{k=1}^{n_x} f_1(k\Delta) + \theta_2 \sum_{k=2}^{n_x} f_1(k\Delta) +$$

$$+ \cdots + \cdots + \theta_{n_x} f_1(n_x\Delta).$$

We now construct the difference between the $\phi$-values ($j \leqslant n_0$):

$$\phi_{x=1}(n_x\Delta) - \phi_{x=0}(n_0\Delta)$$

$$= r \left\{ \sum_{k=0}^{n_x} f_1(k\Delta) - \sum_{k=0}^{n_0} f_0(k\Delta) \right\} + \cdots + \theta_j \left\{ \sum_{k=j}^{n_x} f_1(k\Delta) - \sum_{k=j}^{n_0} f_0(k\Delta) \right\} + \cdots +$$

$$+ \theta_{n_0} \sum_{n_0}^{n_x} f_1(n_0\Delta) - \theta_{n_0} f_0(n_0\Delta) + \theta_{n_0+1} \sum_{k=n_0+1}^{n_k} f_1(k\Delta) + \cdots + \theta_{n_x} f_1(n_x\Delta)$$

$$= r \left\{ \sum_{k=0}^{n_x} f_1(k\Delta) - \sum_{k=0}^{n_0} f_0(k\Delta) \right\} + \cdots +$$

$$+ \theta_j \left\{ \sum_{k=0}^{n_x} f_1(k\Delta) - \sum_{k=0}^{n_0} f_0(k\Delta) \right\} - \theta_j \left\{ \sum_{k=0}^{j-1} f_1(k\Delta) - \sum_{k=0}^{j-1} f_0(k\Delta) \right\} + \cdots +$$

$$+ \theta_{n_0} \left\{ \sum_{k=0}^{n_x} f_1(k\Delta) - \sum_{k=0}^{n_0} f_0(k\Delta) \right\} - \theta_{n_0} \left\{ \sum_{k=0}^{n_0-1} f_1(k\Delta) - \sum_{k=0}^{n_0-1} f_0(k\Delta) \right\} +$$

$$+ \text{positive terms.}$$

Since $\sum_{k=0}^{n} f_1(k\Delta) < \sum_{k=0}^{n} f_0(k\Delta)$, $\sum_{k=0}^{n_x} f_1(k\Delta) - \sum_{k=0}^{n_0} f_0(k\Delta) = 0$, and $\theta_i \geqslant 0$,

the expressions above can be simplified. We obtain:

$$\phi_{x=1}(n_x\Delta) - \phi_{x=0}(n_0\Delta) = 0 + \cdots + 0 - \text{ neg. term } + \cdots + 0 - \text{ neg. term} +$$

$$+ \text{positive terms,}$$

i.e.
$$\phi_{x=1}(n_x\Delta) - \phi_{x=0}(n_0\Delta) > 0.$$

In order to interpret this inequality we introduce some new notations. By $y_\tau$ we denote the $y$-value corresponding to $n_x\Delta$ and by $y_0$ the $y$-value

corresponding to $n_0\Delta$. We have chosen $n_x$ and $n_0$ so that the following equalities hold:

$$F_{x=1}(n_x\Delta) = F_{x=1}(y_x) = \sum_{k=1}^{n_x} f_{x=1}(k\Delta) = \sum_{k=1}^{n_0} f_{x=0}(k\Delta) = F_{x=0}(y_0) = F_{x=0}(n_0\Delta).$$

Further we find:

$$\phi_{x=1}(n_x\Delta) - \phi_{x=0}(n_0\Delta) = \phi_{x=1}(y_x) - \phi_{x=0}(y_0) > 0.$$

If we had chosen $y_x = y_x^* = Y_{x=1}^{\text{Med}}$ and by $y_0^*$ had denoted the value corresponding to $y_x^*$ so that $F_{x=1}(y_x^*) = F_{x=0}(y_0^*)$, we could write

$$\phi_{x=1}(y_x^*) - \phi_{x=0}(y_0^*) > 0.$$

This expression is equivalent to

$$\phi_{x=1}(Y_{x=1}^{\text{Med}}) - \phi_{x=0}(y_0^*) > 0.$$

From this it follows since $\phi$ is non-decreasing that $Y_{x=0}^{\text{Med}} \geqslant y_0^*$ and hence

$$Y_{x=1}^{\text{Med}} - Y_{x=0}^{\text{Med}} < y_x^* - y_0^*$$

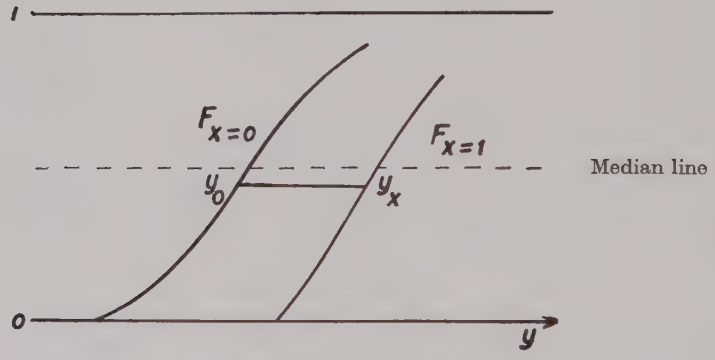



Fig. 13.

Among all the possible differences, $y_x - y_0$ (the horizontal distances between $F_{x=1}$ and $F_{x=0}$ below the median line) at least one is larger than $Y_{x=1}^{\text{Med}} - Y_{x=0}^{\text{Med}}$. Moreover only such $y_x$ (and $y_0$) which can possibly be population medians are relevant; i.e. since $R(y)$ is a non-increasing function, the smallest $y_x$-value allowed satisfies the equation $F_{x=1}(y_x) = \frac{1}{2}\, p_{x=1}$ and the largest $y_x$-value the equation $F_{x=1}(y_x) = \frac{1}{2}$ (and analogically for $y_0$). The largest of the possible $y_x - y_0$ gives a max. value for $Y_{x=1}^{\text{Med}} - Y_{x=0}^{\text{Med}}$.

*Estimation of a min. value for the median difference.*

A min. value solution is also possible as is shown in Fig. 14:

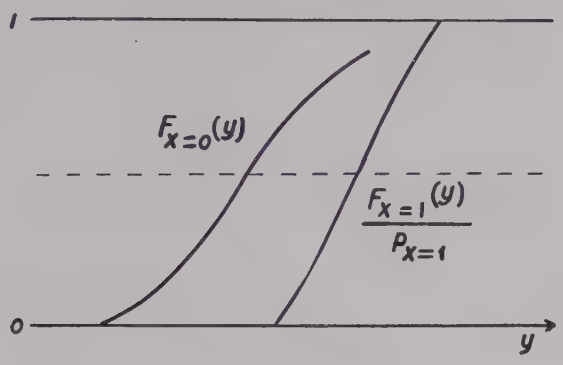



Fig. 14.

A min. value is obtained as the horizontal distance between the two curves along the broken line.

This solution can be sharpened.

**Approach IV *b*.**

*Min. and max. values for the median differences.*

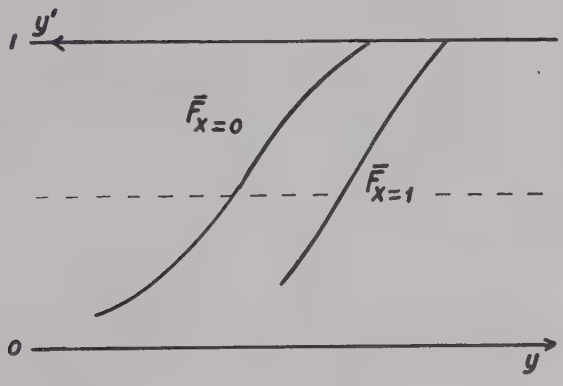



Fig. 15.

It is possible to find a solution along the same lines as app. IV *a*. This is possible if we substitute $F_{x=0}$ by $\bar{F}_{x=1}$ and $F_{x=0}$ by $\bar{F}_{x=1}$ etc. By rotating Fig. 15 the analogy becomes evident (the $y'$-axis will be in the same position as the $y$-axis in Fig. 13). In this way we can also

find a max. value for the median differences, $Y^{\text{Med}}_{x=1} - Y^{\text{Med}}_{x=0}$. A min. value difference between $Y^{\text{Med}}_{x=1} - Y^{\text{Med}}_{x=0}$ is illustrated in Fig. 16.

The broken line between the curves corresponds to the median difference in Fig. 16.

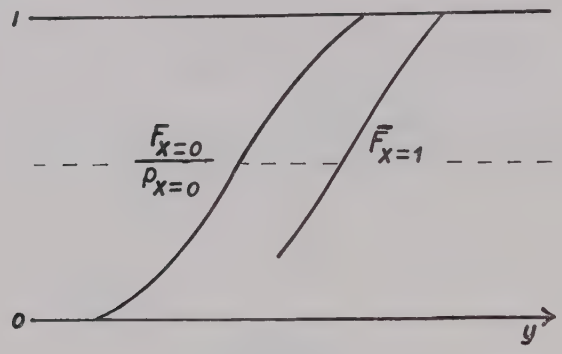



FIG. 16.

**Approach IV *c*.**

*Assumptions:*

We assume that either $R(y)$ is a non-increasing function or a non-decreasing function. We know from the observed material that

$$F_{x=0}(y) \geqslant F_x(y) \quad \text{and} \quad \bar{F}_{x=0}(y) \geqslant \bar{F}_x(y) \quad \text{for all y.}$$

*Max. and min. solutions of the median differences.*

The larger of the max. values obtained in app. IV *a* and app. IV *b* gives a max. estimate of the median differences, and the smaller of the min. values obtained in app. IV *a* and app. IV *b* gives a min. estimate.

**Approach IV *d*.**

*Assumptions:*

$$\phi_{x=0}(y) \geqslant \phi_{x=1}(y); \quad R(y) = R_{x=0}(y) = R_{x=1}(y).$$

*Notation:*

We divide the material into two groups and denote by $y$ the values for which $R(y) > 0$ and denote by $y'$ the values for which $R(y') = 0$. We introduce $\pi_{x=0}(y')$ and $\pi_{x=1}(y')$, where $\pi_{x=0}(y')$ corresponds to the relative frequency of values for which $R_{x=0}(y') = 0$; and analogically, $\pi_{x=1}(y')$ for which $R_{x=1}(y') = 0$.

The following relations can be derived:

The relative frequency adds up to 1 in the two sub-populations where $x=0$ and $x=1$. The sums shall be taken for all $y$- and $y'$-values.

$$\sum \frac{f_{x=0}(y)}{R_{x=0}(y)} + \sum \pi_{x=0}(y') = 1 \tag{63}$$

and

$$\sum \frac{f_{x=1}(y)}{R_{x=1}(y)} + \sum \pi_{x=1}(y') = 1. \tag{64}$$

For the frequency of individuals in the remaining material it holds:

$$0 < R(y) \leqslant 1 \quad \text{for every } y. \tag{65}$$

Further the following inequalities must be valid:

$$\pi_{x=0}(y') \geqslant 0 \quad \text{for every } y = y' \tag{66}$$

and

$$\pi_{x=1}(y') \geqslant 0 \quad \text{for every } y = y'. \tag{67}$$

Since $\phi_{x=0}(y^*) \geqslant \phi_{x=1}(y^*)$ for every $y^*$:

$$\sum \frac{1}{R(y)} (f_{x=0}(y) - f_{x=1}(y)) + \sum (\pi_{x=0}(y') - \pi_{x=1}(y')) \geqslant 0 \text{ for every } y^*. \tag{68}$$

The sums shall be taken over all $y \leqslant y^*$ and $y' \leqslant y^*$.

We want to estimate:

$$D = \sum y \left\{ \frac{f_{x=1}(y)}{R_{x=1}(y)} - \frac{f_{x=0}(y)}{R_{x=0}(y)} \right\} + \sum y' \left\{ \pi_{x=1}(y') - \pi_{x=0}(y') \right\}. \tag{69}$$

A linear-programming approach is possible since the relations are linear in $1/(R(y))$, $\pi_{x=0}(y')$, and $\pi_{x=1}(y')$[1], while the values of $f_{x=0}(y)$ and $f_{x=1}(y)$ are known from the material. The linear programming gives us a min. and a max. estimator (finite or infinite) for $D$. For references concerning linear programming see e.g. Bellman, 1957.

**Approach IV *e***

*Assumptions:* $\phi_{x=0}(y) \leqslant \phi_{x=1}(y)$ for every $y$,

$$R(y) = R_{x=0}(y) = R_{x=1}(y) \text{ for every } y.$$

---

[1] The linear-programming approach is made possible after a classification of the $y$-values in a finite number of classes is performed. We assume that the values of $R(y)$, $\pi_{x=0}(y')$, $\pi_{x=1}(y')$ do not vary within classes.

*Estimation of a difference between population means.*

We want to estimate $D$, as in (69). The same relations as those in app. IV $d$ are valid, except (68). Since $\phi_{x=0}(y^*) \leqslant \phi_{x=1}(y^*)$ for every $y^*$:

$$\sum \frac{1}{R(y)} (f_{x=0}(y) - f_{x=1}(y)) + \sum (\pi_{x=0}(y') - \pi_{x=1}(y')) \leqslant 0 \quad \text{for every } y^*. \tag{70}$$

The sums shall be taken over all $y \leqslant y^*$ and $y' \leqslant y^*$.

Linear programming can be applied to the relations as in app. IV $d$. A max. and min. value for $D$ can be estimated in this way.

**Approach IV $f$**

*Assumptions:* Either $\phi_{x=0}(y) \leqslant \phi_{x=1}(y)$ is valid for every $y$,
or $\phi_{x=0}(y) \geqslant \phi_{x=1}(y)$ is valid for every $y$, and
$R = R_{x=0} = R_{x=1}$.

*Estimation of a difference between population means.*

According to the assumptions, either (68) or (70) is valid. Moreover the relations common to app. IV $d$ and app. IV $e$ are still valid. The larger of the $D_{\max}$-values obtained in app. IV $d$ and app. IV $e$ is the relevant $D_{\max}$-value, while the smaller of the $D_{\min}$-values obtained in app. IV $d$ and app. IV $e$ is the relevant $D_{\min}$-value.

**Approach IV $g$**

*Assumptions:* $R(y) = R_{x=0}(y) = R_{x=1}(y)$,
$R(y)$ is a non-increasing function of $y$,
$\phi_{x=0}(y) \geqslant \phi_{x=1}(y)$ for every $y$.

*Estimation of the difference of population means.*

The same relations as those in app. IV $d$ are valid. Since $R(y)$ is non-increasing we can add the assumption:

$$R(y + \Delta) - R(y) \leqslant 0 \quad \text{for every } y \quad (\Delta > 0). \tag{71}$$

A linear-programming solution is still possible, giving max. and min. values for $D$.

*Remark:* From the relations (63), (64), (65), (66), (67), and (71) a solution by linear programming can be derived for app. IV $a$, giving max. and min. values for $D$ as defined in relation (69).

Analogous arguments are also relevant to approaches IV $b$ and $c$.

## 5.8. Approach V

*Arrow scheme:* See p. 75.

One or more of the following *assumptions* are valid:

(a) $R_{x=1}(y)$ is a non-decreasing function of $y$.

(b) $R_{x=0}(y)$ is a non-decreasing function of $y$.

(c) $R_{x=1}(y)$ is a non-increasing function of $y$.

(d) $R_{x=0}(y)$ is a non-increasing function of $y$.

(e) $R_{x=1}(y) \geqslant R_{x=0}(y)$ for every $y$.

(f) $R_{x=1}(y) \leqslant R_{x=0}(y)$ for every $y$.

(g) $\phi_{x=0}(y) - \phi_{x=1}(y) \geqslant 0$ for every $y$.

(h) $\phi_{x=0}(y) - \phi_{x=1}(y) \leqslant 0$ for every $y$.

### Approach V *a*

*Linear-programming approaches.*

We want to estimate the difference $D$ between population means (see eq. (69)). The following relations are basic, i.e. independent of which assumptions are added from the above list: (63), (64), (66), (67), (72) and (73).

$$\begin{aligned} 0 < R_{x=0}(y) \leqslant 1 \text{ for every } y. \\ 0 < R_{x=1}(y) \leqslant 1 \text{ for every } y. \end{aligned} \tag{72}$$

The assumptions included in the list correspond to the following relations:

(a) $R_{x=1}(y+\Delta) - R_{x=1}(y) \geqslant 0$ for every $y$ $(\Delta > 0)$. (73)

(b) $R_{x=0}(y+\Delta) - R_{x=0}(y) \geqslant 0$ for every $y$ $(\Delta > 0)$. (74)

(c) $R_{x=1}(y+\Delta) - R_{x=0}(y) \leqslant 0$ for every $y$ $(\Delta > 0)$. (75)

(d) $R_{x=0}(y+\Delta) - R_{x=0}(y) \leqslant 0$ for every $y$ $(\Delta > 0)$. (76)

(e) $R_{x=1}(y) \geqslant R_{x=0}(y)$ for every $y$. (77)

(f) $R_{x=1}(y) \leqslant R_{x=0}(y)$ for every $y$. (78)

(g) $\sum \{f_{x=0}(y)/R_{x=0}(y) - f_{x=1}(y)/R_{x=1}(y)\} +$
$+ \sum \{\pi_{x=0}(y') - \pi_{x=1}(y')\} \geqslant 0$ for every $y^*$. (79)

(h) The expression on the left hand side in (g) is $\leqslant 0$ for every $y^*$. (80)

Linear programming is possible for every combination of the above assumptions. In this way we obtain max. and min. values for $D$. Some of these values are irrelevant since the extreme values are infinitely large. (Cf. app. IV *d*).

### Approaches V *b*, *c*, *d* and e

*Median approaches.*

On p. 86 it was illustrated that a finite median solution is possible if $p_x = p_{x=1} > \frac{1}{2}$ and $p_0 = p_{x=0} > \frac{1}{2}$.

### Approach V *b*

*Assumptions:* $R_{x=0}(y)$ non-increasing,
$R_{x=1}(y)$ non-increasing,
$R_{x=1}(y) \leqslant R_{x=0}(y)$,
$F_{x=1}(y) \leqslant F_{x=0}(y)$.

*Solution:* A max. solution analogous to the one given in app. IV *a* can be given.

### Approach V *c*

*Assumptions:* $R_{x=0}(y)$ non-decreasing,
$R_{x=1}(y)$ non-decreasing,
$R_{x=1}(y) \geqslant R_{x=0}(y)$,
$\overline{F}_{x=1}(y) \leqslant \overline{F}_{x=0}(y)$.

*Solution:* A max. solution analogous to the one given in app. IV *b* can be given.

### Approach V *d*

*Assumptions:* $R_{x=0}$ is a non-increasing function,
$R_{x=1}$ is a non-increasing function.

*Solution:* Max. and min. solutions of the median differences are illustrated in a simple way.

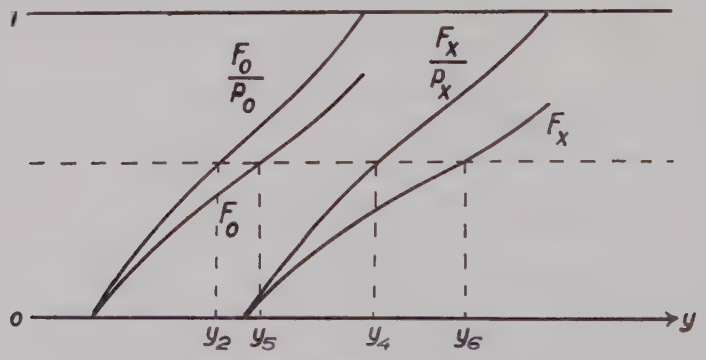



Fig. 17.

We obtain: Max. median difference $y_6 - y_2$; and min. median difference $y_4 - y_5$. Note that a finite max. solution cannot be found for $D$.

## Approach V *e*

*Assumptions:* $R_{x=0}$ is a non-decreasing function,
$R_{x=1}$ is a non-decreasing function.

*Solutions:* (see Fig. 18):

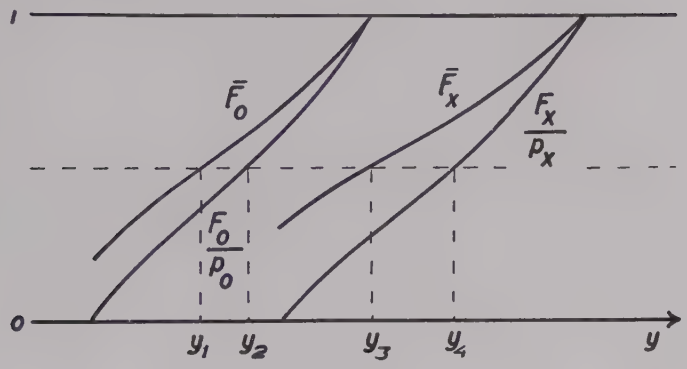



FIG. 18.

Max. value of the median differences in the population: $y_4-y_1$ Min. value: $y_3-y_2$.

CHAPTER 6

# SELECTION BIAS DUE TO COMMON CAUSAL FACTOR

## 6.0. Introduction

The aim of this chapter is to present current methods suitable for the statistical treatment of problems arising from a hidden common causal factor ($w$). We start by some examples on regression problems, i.e. problems where the genuine effect of a certain explanatory variable ($x$) on a dependent variable ($y$) is studied (cf. 1.2).

## 6.1. Arrow scheme

The following arrow scheme is valid in this particular type of problem (cf. 3.63):

We wish to estimate the effect of $x$ on $y$. Here, as earlier, we shall denote a certain factor, or possibly a complex of factors called the *common causal factor(s)* by $w$. We use the notation even if $w$ can only be suspected of having an effect on $x$ and $y$. Observe that $w$ is farthest back in the causal chain. Thus $w$ constitutes an underlying causal factor (cf. 2.33).

## 6.2. Investigations in which a selective factor $w$ is suspected

A series of investigations from different fields is set out below, in which the object was to study the effect of an explanatory variable $x$ on a dependent variable $y$ (cf. comments, 6.21): Exs. 1–4 are derived from medicine, Exs. 5–7 from odontology, Ex. 8 from criminology, Exs. 9–10 from demography, Ex. 11 from pedagogy, Exs. 12–13 from economics.

All the above examples follow the arrow scheme which has been given above. The factor $w$ can consequently be suspected of having an effect on both $x$ and $y$. Neither does there appear to be any uncertainty of the causal direction between $x$ and $y$.

*Comment* 1. The information concerning the observation material is somewhat schematic. One reason for giving it here is to enable the reader to

TABLE. *Survey of investigations in which a common causal factor, w, can be indicated.*

| $x$ Studied explanatory factor (groups which are compared) | $y$ Studied dependent variable | $w$ Common causal factor | Source Cf. the reference list |
|---|---|---|---|
| Ex. 1, pregnant and nonpregnant cases of lung tuberculosis | development of tuberculosis during pregnancy | psychic factors | Hedvall |
| Ex. 2, children of mothers with and without rubella during pregnancy | frequency of malformation or body weight | risk of infection for mother | Lundström |
| Ex. 3, breast milk and cows milk for newborn infants | frequency of infection | risk of infection (via mother) | Vahlquist & Melander |
| Ex. 4, smoking | cancer | degree of urbanisation and sex | Wallis & Roberts |
| Ex. 5, rubber and acrylic as prosthetic materials | palatal ulcer denture sore mouth | age of dentures and age of patient | Nyquist |
| Ex. 6, fluoride content in water | caries | salts in water | Abramson |
| Ex. 7, tartar | caries | oral hygiene and yeasts in oral cavity | Welander |
| Ex. 8, living conditions | criminality | parental economy | |
| Ex. 9, freeholders and farm workers | number of children | age | v. Hofsten |
| Ex. 10, income level or unemployment | marriage rate or birth rate | other trends | Hyrenius |
| Ex. 11, school education | intelligence quotient | heredity and parents' education | Husén Gesell |
| Ex. 12, prices of liquor | liquor consumption | income | Malmquist |
| Ex. 13, income | food comsumption | size of family | Wold (1940) |

6.21. *Comments on the observation data used in the examples.*

Ex. 1): Follow-up study. A group of pregnant women and another group of non-pregnant ones with the same age distribution and, as it seems, in the same initial stage of tuberculosis are compared. After one or two years the development of tuberculosis is studied. (See Hedvall E. *Pregnancy and tuberculosis,* 1953.)

Ex. 2): Women in Sweden who had rubella in 1951–1952 during pregnancy and a sample of pregnant women not having rubella have been selected. The children of the respective groups are compared with respect to malformations etc. (at birth and at the age of about 3 years).

Ex. 3): Data concerning all children born during 1953–54 within a region in northern Sweden.
Ex. 4): Cf. section 4.23.
Ex. 5): Cf. section 1.21 example c.
Ex. 6): Cf. section 1.23 example d.
Ex. 7): Cross-sectional investigation based on a sample consisting of every third recruit from some recruitment areas in Sweden 1942.
Ex. 8): Cross-sectional data taken in Sweden, where different types of neighborhoods were compared with respect to the degree of criminality.
Ex. 9): Cross-sectional data from a census concerning different families in Sweden.
Ex. 10): Census data for Sweden in relation to different economic indicators such as the number of unemployed and average income.
Ex. 11): Undergraduate recruits in the age group 20–21 compared with others with respect to intelligence according to a test.
Ex. 12): Cf. section 1.21 ex. b.
Ex. 13): Data concerning annual income and food consumption for a sample group in Sweden.

understand more easily how the values of the variables $x$ and $y$ have arisen. It should further be emphasized, that realistic and evident judgements concerning the causal relationships between $x$ and $y$ are possible only if one knows how the material was collected.

*Comment* 2. We disregard possible "nonresponse" or "incomplete data" present in the above problems.

*Comment* 3. Factor $w$ can often be represented as a "general" underlying factor; here belong e.g. age (Exs. 5 and 9), parental economy (Ex. 8) and education (Ex. 11), degree of urbanisation (Ex. 4), sex (Ex. 4). Sometimes $w$ is a more specific factor as for example infection risk in (Ex. 3), yeasts in (Ex. 7) and age of dentures in (Ex. 5). Factors of type $w$ appear in different forms and vary from investigation to investigation.

*Comment* 4. The arrow scheme


should not be confused with the arrow scheme on p. 98.

The aim of the above examples is to study the relation $x \rightarrow -y$. Our main interest is:

(i) To estimate how $y$ changes when $x$ changes (as in Ex. 6: How much does caries frequency diminish, if the fluoride content of the water is increased? Or as in Ex. 2: How much greater is the frequency of malformation among children of mothers suffering from rubella during pregnancy, than among other children?), or more simply,

(ii) Has $x$ a positive or a negative effect on $y$? Instances are Ex. 3: Does feeding with breastmilk bring lower (or higher) frequency of infection in

children?, or Ex. 5: Do rubber prostheses give rise to lower (or higher) frequency of palatal ulcer than acrylic dentures?, or Ex. 2: Is the frequency of malformation greater or less than among control children?

When we wish to study these problems on observation material which is composed only of $(x, y)$ data for different individuals (or years etc.) we meet difficulties of interpretation in those cases where the existence of a $w$-factor may be suspected, e.g. see the examples a, b, c in section 1.2.

## 6.3. Current methods of treating this selection

In the present section, we shall survey different methods, referred to in the literature, and useful, when handling this selection mechanism. Only in exceptional cases have I tried to give original references[1]—most of the methods referred to here are in general use.

### 6.31. *Experimental versus nonexperimental approach*

The most effective way is, naturally, to apply a randomised experiment if such is possible.

Some examples are given in the adjoining table to show practically how the experimental and the nonexperimental procedure differs from each other.

Table.

| Problem complex | Experimental procedure | Nonexperimental |
|---|---|---|
| Ex. 1. Concerning disease (rise and development of cancer) | animal experiments with cancer stimulating substances | clinical investigations and enquiries (e. g. lung-cancer and smoking) |
| Ex. 2. Concerning the pathogenesis of caries | laboratory experiments on solubility in vitro | observations on patient material. Chemical analysis of saliva etc. in people with high and low caries activity |
| Ex. 3. Concerning laws of heredity | animal and plant experiment | field investigations of a population |
| Ex. 4. Concerning drugs and therapies | control (placebo group) and treated group of persons or animals | observations on patients with co-operation of practising doctors |

[1] In order to demonstrate that these methods have been known for a long time I repeatedly resorted to Yule's elementary textbook, partly because the first Edition had already been published at about 1910 and partly because it has a very wide circulation and has certainly exerted a strong influence on "Applied Statistics".

A development from the nonexperimental to experimental approach is rather usual, as is seen above in Exs. 1, 2 and 4. With regard to its interpretation, the experiment appears as an ideal for elucidating questions of the size of an effect (cf. 2.23).

In many spheres however, a direct experiment cannot be undertaken; in medicine one is often compelled to turn to animal experiments, in accordance with medical practice and the fact that experiments on human material may not be obtainable. For this and other reasons one often finds that experimental and nonexperimental procedures can occur side by side.

In sociology and social psychology other obstacles of a practical nature are met (see Chapin: *Experimental Designs in Sociology*, 1947).

In child psychology, social norms of upbringing present important obstacles, and in economics random experiments are impossible, for political reasons among others.

### 6.32. "*Nature's experiment*"

A good approximation to randomised experiment can sometimes be obtained by so-called "nature's experiment", e.g. (*a*) sex differences are able to be tested, since that component which decides the sex genetically can be assumed to act by chance, when disregarding child mortality, and (*b*) the importance of the parental influence on the child's development (emotional, intelligence, etc.) studied on uni-ovular twins, separated soon after birth, with foster parents from different environments (see Newman, Freeman & Holzinger, 1936).

In the above examples it is difficult to conceive of a $w$-factor.

Many other examples could certainly be given, but one can, however, only exceptionally hope for such an easy solution. As a rule, one can directly point out a definite factor $w$, suspected to have an effect on both $x$ and $y$. The next step therefore is to try and eliminate such $w$-factors.

### 6.33. *Regression analysis and related methods*

The genuine correlation between $x$ and $y$ or the genuine effect of $x$ on $y$, may in principle be carried out by multiple correlation analysis, regression analysis, standard weighting, or variance analysis. Among those factors which can be suspected of having an effect on both $x$ and $y$, as many as possible should be included in the analysis as secondary explanatory variables (cf. section 2.52).

*Illustration:* Analysis of demand, market prices for a number of years.

Dependent variable ($y$): consumption per capita of some agricultural product.

Studied explanatory variable ($x$): price level.

Secondary explanatory variables:

Consumers' average income
Price of competing goods
Advertising for goods
Time trend (including, changes of taste etc.)
Professional distribution, worker, official
Age distribution
Income distribution.

Even if it appears that all relevant factors are included in the analysis, one cannot however, entirely free oneself from the suspicion that one, or more, hidden, biasing $w$-factor(s) still remain.

Arguments have been brought forward proposing that the analysis should nevertheless be acceptable in every case as a first approximation:

($a$) *High total correlation* implies that the most relevant factors are included.

($b$) *Small residual* constitutes a closely related argument to this.

($c$) If $w$-factors are poorly correlated with $x$, the bias will be negligible.

($d$) A co-ordination of arguments in ($b$) and ($c$) can be summarized in the Proximity Theorem (stated by Wold & Juréen, 1952, page 37; cf. the quotation below):

> It is obvious that the regressions coefficients are unbiased *either* if the disturbances factors are negligible (so that the residuals are small), or if they are uncorrelated with the regressors. The proximity theorem shows that the two conditions strengthen each other to the effect that if they are fulfilled up to magnitudes that are small of the *first* order that bias of the regressions coefficients will be small of the *second* order.
>
> *Illustration.* Supposing that $y$ is linearly dependent upon $x$, let the theoretical and empirical regression relations be written
>
> $$y = \beta x + z^* \quad y = bx + z \quad \text{with} \quad Ex = Ey = 0.$$
>
> In symbols easily understood, if then $\sigma(z^*) = 1/5\ \sigma(x)$ and $r(x, z^*) = 1/5$, we obtain
>
> $$(1) \qquad E(b) = \frac{E(xy)}{E(x^2)} = \beta + \frac{E(xz^*)}{E(x^2)} = \beta + r(x, z^*) \cdot \frac{\sigma(z^*)}{\sigma(x)} = \beta + \frac{1}{25}.$$
>
> The bias in the empirical coefficient thus is only about 0,04."

Cf. Wold & Faxér, 1957, in which a new and more general version of this argument has been given.

(*e*) If one has a good knowledge of the field of application, and can consider the most important factors, the concealed factors possibly remaining should produce only a negligible bias.

(*f*) *A strong effect* can seldom be brought back on a disturbing variable.

(*g*) If repeated investigations among other populations give the same result (about the same coefficients) it is reasonable to accept the estimate of regression.

### 6.34. *Matching*

The aim of matching is to eliminate disturbing factors. In the analysis of a material—or already in planning the investigation—pairs of subjects having the same values of the underlying (secondary) factors, but differing with respect to $x$-values, are matched. Then $\bar{y}$ for $x = 1$ and $\bar{y}$ for $x = 0$ is compared.

The following example of matching is quoted from Chapin (1947, p. 29):[1]

"Control of social conditions is obtained not by manipulating people or by exerting any physical force on persons. The control is obtained by *selecting* for observation two groups of like individuals, for example, individuals of the same income bracket, the same occupational class, the same chronological age, the same size of family, the same intelligence quotient, etc. (by matching on these attributes). The one group, called the experimental group, is given treatment, or receives some social program, or is subjected to some assumed and uncontrolled natural force ($F$) in the environment, while the other group, called the control group, is denied this treatment, program, of force ($F$)."

As an example of successful matching, investigations on twins should be mentioned; a large number of genetic factors can be controlled by matching (cf. Newman, Freeman & Holzinger, 1936).

Lundström's rubella investigation provides another example. Because each rubella child was matched with a control child from the same hospital and born almost at the same time, it was nearly always the same doctor who judged the birth status of both children, so that an important cause of variation is eliminated in the comparison.

In matching, consideration can also be given to the subjective impression that certain individuals seem to be similar in all respects except with regard to $x$ and $y$. In this way, such factors as e.g. health and psyche which are otherwise difficult to express quantitatively can also be checked. The matching technique can evidently be made more flexible than regression analysis.

[1] Objections have been raised against the concepts "experimental group" and "experimental design" used by Chapin. Note that the material is nonexperimental.

A disadvantage with matching can be the large loss of material which accompanies this type of paired matching, recommended by Chapin. It is true that the loss of material can be partly avoided because matching groups can comprise more than a pair (cf. Nyquist) but is seen nevertheless to constitute a difficult obstacle for a more general application of this technique. This remark is nevertheless valid, if matching can only be accomplished afterwards. This disadvantage is invalid if, matching can be done at the stage of collecting material (prospective investigations), as was the case with the rubella investigation.

### 6.35. *Dynamic approach*

With dynamic investigations, i.e. where changes with time are studied, one can often—through choice of suitable dependent, and possibly also explanatory, variables—eliminate in every case certain types of disturbing concealed variables, e.g. those, $w$, which certainly affected the initial values for $x$ and $y$, but no longer affect their changes in the following course. The changes in $x$ and $y$, from one point of time to another, may conveniently be studied. It has long been known that greater precision is achieved by estimating of the effect of $x$ on $y$ through an analysis of processes (see Yule's elementary textbook, 1911); on the other hand it is difficult to find a reference in the literature recommending dynamic approaches, on account of the fact that they can also directly reduce bias introduced by disturbing concealed variables.

The following statement by Lundberg indicates how important the *dynamic approach* and *matching* is in sociology (Lundberg, 1942, p. 65):

"In social experiments especially, it is usually necessary to experiment on a single group, and to draw our conclusions from a comparison of the significant facts *before* and *after* a new factor is introduced.

This gives rise to serious complications owing to the lack of control of important conditions during the whole period of the experiment. But if we cannot *inject* or *eliminate* external conditions which may influence the outcome of an experiment, may not the same result be attained by *measuring the effect* of these influences? In this way we can allow for the influence of each variable in the gross results, and conclude that any residuary difference in social conditions before and after a new factor was introduced must be due to this factor. If we keep a careful record of *all* significant social conditions, for example, before a law is enacted, and keep *strictly comparable* records after its enactment, may we not draw conclusions of some validity regarding the results of the experiment? If we cannot take into consideration and make allowance for *all* the conditions that influence an event over a period of time, which is probably impossible, we may be able to get a "first approximation" to the truth."

The estimator obtained in a dynamic approach is, however, not always the same as that obtained in a cross-sectional approach. An immediate rise in income, e.g., may have another effect on saving than a slow income rise of the same amount.

### 6.36. *Bias in an unimportant direction*

One sometimes knows in advance whether $w$ has a positive or negative effect on $x$ and $y$ respectively (cf. 1.31). If we know that $r_{xw}>0$ and $r_{yw.x}>0$ and we find that $b_{yx}>0$, then we can derive that the regression coefficient $\beta_{yx.w}>E(b_{yx})$. In this case $b_{yx}$ as a rule is less than $\beta_{yx.w}$.

### 6.37. *Instrumental variable, z, with strong effect on x, but independent of w*

The approach by means of instrumental variables was originally introduced by Reiersøl (1941, 1945) as being suitable in applications to economic time series data. Reiersøl's approach is applicable to selection problems due to hidden factor(s). In connection with such selection problems it can be regarded as an application of Wright's path analysis.[1]

We are going to illustrate how to use an instrumental variable in estimating the path coefficient, $P_{yx}$. We start from the arrow scheme below, where we have assumed that $\varrho_{zw}=0$ and $\varrho_{yz.x}=0$.

*Arrow scheme:*

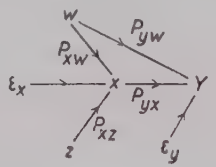



Fig. 19.

*Legend:* Observed variables: $x$, $y$, $z$ Observed correlations: $r_{yx}$, $r_{xz}$, $r_{yz}$. $P_{yz}=0$, i.e. $z$ has no other effect on $y$ than via $x$. $P_{zw}=0$.

All relations between $x$, $y$, $z$ and $w$ are assumed to be linear (cf. 2.36). By means of the known correlations we obtain the following equations:

$$r_{yx}=P_{yw}P_{xw}+P_{yx},$$
$$r_{xz}=P_{xz},$$
$$r_{yz}=P_{xz}P_{yx}.$$

[1] Wright, however, has not explicitly stated that his path analysis could provide a solution for this problem, but this may be due to the fact that he has been primarily interested in genetic problems. He has *realized*, however, that "path analysis" has wide applications.

From this it follows

$$\frac{r_{yz}}{r_{xz}} = P_{yx},$$

If our aim had been to estimate the regression coefficient $\beta_{yx.wz}$ (or $\beta_{yx.w}$ which must be identical if the model is correctly specified), we simply would obtain (cf. 2.36) that

$$P_{yx} = \beta_{yx.w}\frac{\sigma_x}{\sigma_y} = \beta_{yx.wz}\frac{\sigma_x}{\sigma_y},$$

i.e. $\beta_{yx.wz}$ can be estimated as: $\dfrac{r_{xy}\,\sigma_y}{r_{xz}\,\sigma_x}$.

A variant of Reiersøl's approach would give the same solution. In fact, his approach is based on linear confluence analysis. In such analysis an exact linear relation, $y' = a + \sum \beta_i x_i'$, between the systematic components $y'$ and $x'$ is assumed; and $y$ and $x_i$ are observed values, for which the following relations hold: $y = y' + \varepsilon_y$ and $x_i = x_i' + \varepsilon_{xi}$, where $\varepsilon_y$ and $\varepsilon_{xi}$ are random disturbances. It would seem however that Reiersøl is only peripherically interested in selection problems.

His works have been commented on by Geary (1949), by Durbin (1954), and by Malinvaud (1957); a summary is found in Geary (1953). Some econometrical applications are found in Reiersøl (1945) and Geary (1949), but these should be regarded as purely numerical illustrations of the method.

The wide applicability of Wright's method mentioned above seems not to have been observed.

CHAPTER 7

# NEW APPROACHES IN TREATING BIAS DUE TO A COMMON CAUSAL FACTOR

## 7.1. Introduction

In this chapter certain models are presented. The aim is to estimate the genuine effect of $x$ on $y$, i.e. the effect remaining after selective factor(s) have been considered. The presentation follows the arrangement given in section 1.3. The same problem as in Chap. 6 is treated.

In the following survey every model is characterized with respect to assumptions and relevant approaches.

*Model I:* Ordinary regression analysis with one explaining variable. The causal direction is evident.

*Model II:* Ordinary regression analysis with two explaining variables, $x$ and $z$. The causal direction is evident, $x$ and $z$ both have an effect on $y$.

*Model III:* Estimation by means of an instrumental variable, $z$. The causal relation is like that in model II. We further assume that $z$ is zero-correlated with $w$ and that $\varrho_{zy.x} = 0$. Model I and II can be treated by the ordinary regression approach. Model III can be treated by the methods given by Wright and Reiersøl (cf. 6.37). Model IV–VII are new approaches, capable of giving a min. or max. estimator for the genuine effect of $x$ on $y$.

*Model IV:* (i) Causal relation as in model II but $\varrho_{zy.x} \neq 0$ is admitted. (ii) $z$ and $x$ are dichotomous assuming the values 0 and 1. A necessarry but not sufficient condition for $x = 1$ is that $z = 1$.

*Model V:* (i) The same assumption as in model IV (i). Instead of IV (ii) we assume that $r_{xz} > 0$. (ii) $z$ and $w$ are assumed to have no interaction-effect on $y$.

*Model VI:* The same assumptions as in model V. Two instrumental variables are used. Interactions of higher order do not exist. Some special assumptions concerning interactions. $z$ and $w$ have interactions on $x$ and $y$.

*Model VII:* (i) The same assumptions as in model V. (ii) The residual factor $\varepsilon$ is independent of $x$ as well as of $z$ and $w$.

## 7.2. General remarks on the presentation

(a) We determine that the observed group of subjects can be regarded as a random sample from or is the very same as the target population.

(b) All samples are assumed to be large.

(c) We deal with the model as if it were correctly specified.

(d) The aim is to estimate one or more parameters included in the model.

(e) *Notations:* The primary effects are denoted by $\beta$, the interactions of first order by $\gamma$, and the interactions of second order by $\delta$.

In general we write

$$y = \alpha + \text{the primary effect component} + \text{the first order interactions} +$$
$$+ \text{the second order interactions} + \cdots + \cdots + \text{residual factor.}$$

The primary effect component we denote: $\beta_x x + \beta_z z + \sum \beta_{w_i} w_i + \beta_\varepsilon \varepsilon$.

The first order interactions we denote:

$$\gamma_{xz} x z + \sum \gamma_{xw_i} x w_i + \sum \gamma_{zw_i} z w_i + \sum_{i} \sum_{j>i} \gamma_{w_i w_j} w_i w_j + \sum \gamma_{x\varepsilon} x \varepsilon + \sum \gamma_{w_i \varepsilon} w_i \varepsilon.$$

(f) *Material:* $y$, $x$, and $z$ are observed for every subject (complete data) but $w$ and $\varepsilon$ are not observed for any.

(g) The correlation $\varrho$ refers to the "target population".

(h) *Aim of the estimation:* In general we want to estimate the effect of $x$ on $y$ after taking $w$ into account (cf. 1.1). The simplest way of handling $w$ would be to estimate $\beta_x$ in the linear relation:

$$y = a + \beta_x x + \sum \beta_{w_i} w_i + \varepsilon.$$

This is the ordinary multiple regression approach.

The problem is more complicated if the following relation is valid:

$$y = \alpha + \beta x + \sum \beta_{w_i} w_i + \sum \gamma_{xw_i} x w_i + \varepsilon.$$

Here it is evident that the effect of $x$ varies with $w_i$; the effect should be:

$$\beta_x + \gamma_{xw_i} w_i$$

If $w_i$ are hidden variables we must weaken our demands because it is not possible to estimate every $\gamma_{xw_i}$.

For practical reasons it is natural to be satisfied with estimating an average effect of $x$ on $y$. This average estimate might be obtained e.g. by weighting with respect to the different values of $w$. This can be illustrated by the standard weighting technique:

| | Sub-groups | | | Weights | Number of subjects | |
|---|---|---|---|---|---|---|
| | $x=0$ | $x=1$ | | | $x=0$ | $x=1$ |
| $w=w'$ | $\overline{Y}'_0$ | $\overline{Y}'_x$ | $D' = \overline{Y}'_x - \overline{Y}'_0$ | $v'$ | $N'_0$ | $N'_x$ |
| $w=w''$ | $\overline{Y}''_0$ | $\overline{Y}''_x$ | $D'' = \overline{Y}''_x - \overline{Y}''_0$ | $v''$ | $N''_0$ | $N''_x$ |
| $w=w^{(i)}$ | $\overline{Y}^{(i)}_0$ | $\overline{Y}^{(i)}_x$ | $D^{(i)} = \overline{Y}^{(i)}_x - \overline{Y}^{(i)}_0$ | $v^{(i)}$ | $N^{(i)}_0$ | $N^{(i)}_x$ |

We attempt to estimate $\bar{D}$:

$$\bar{D} = \frac{v' D' + v'' D'' + v^{(i)} D^{(i)} + \cdots}{v' + v'' + v^{(i)} + \cdots}.$$

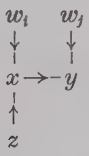
The weights $v'$, $v''$, $v^{(i)}$ etc. can now be chosen in many ways. The aim of the estimation is sometimes fixed in such a way that only a certain weighting system is relevant. More often the problem is vaguely formulated and many weighting systems would be acceptable.

In the models specified below we have assumed that the genuine $x$ effect on $y$ is independent of $w_i$. Hence that weighting problem can be ignored.

## 7.3. Models and approaches

### Model I

*Arrow scheme:*




*Assumptions:*

(1) There is no selective factor $z$ or $w$ having an effect on $x$ as well as on $y$. (Cf. the arrow scheme.)

(2) The dependent variable, $y$, can be expressed as a linear function of $x$, $z$, and $k$ factors, $w_1$, $w_2$, ..., $w_i$, ..., $w_k$, i.e. the mathematical model is:

$$y = \alpha + \beta_x x + \beta_z z + \sum \beta_{w_i} w_i.$$

According to assumption (1) the following correlations must be equal to 0, i.e.:

$$\varrho_{w_i x.yzw_1 \ldots w_{i-1} w_{i+1} \ldots w_k} \neq 0 \text{ for all } w_i \text{ with } \beta_{w_i} = 0,$$

$$\varrho_{zx.yw_1 \ldots w_i \ldots w_k} \neq 0 \text{ and } \beta_z = 0,$$

$$\varrho_{xw_j.zw_1 \ldots w_{j-1} w_{j+1} \ldots w_k} = 0 \text{ for all } w_j \text{ with } \beta_{w_j} \neq 0.$$

*Aim:* Estimation of $\beta_x$.

*Material:* $x$ and $y$ are observed but $z$ and $w_i$ are not observed.

*Solutions:*

(1) By regression analysis with one explaining variable, a consistent estimator of $\beta_x$ denoted $b_x^0$ is obtained

$$b_x^0 = r_{xy} \frac{s_y}{s_x}, \qquad (81)$$

where $r_{xy}$, $s_x$, and $s_y$ can be calculated from the material.

(2) If $x$ is dichotomous, we construct the difference between the mean of $y$ for $x = 1$ and $x = 0$, i.e.

$$\bar{y}_{x=1} - \bar{y}_{x=0}.$$

Solution (2) agrees with solution (1).

**Model II**

*Assumptions:* The same ones as in model I, but here we allow $z$ to have an effect on $x$ and on $y$.

*Aim:* Estimation of $\beta_x$.

*Material:* For every subject $x$, $y$ and $z$ are observed.

*Solution:* A solution can be reached using multiple regression analysis (with two explaining variables). Our estimator $b_{yx.z}$ is calculated according to the least square principle.

$$b_{yx.z} = \frac{s_y}{s_x} \frac{r_{yx} - r_{yz} r_{xz}}{1 - r_{xz}^2}. \tag{82}$$

We also know that $Eb_{yx.z} = \beta_x$ (disregarding a magnitude 0 $(1/n)$).

**Model III**

This model illustrates an approach where an instrumental variable is used.

*Assumptions:*

(1) $w$ has an effect on $x$ as well as on $y$.
$w$ is an abbreviated notation for all the factors $w_1 \ldots w_i \ldots w_k$.

(2) $z$ has effect on $x$, so that $r_{zx} > 0$.

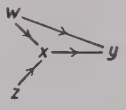
(3) $z$ has no effect on $y$ other than the effect via $x$, i.e. $\varrho_{zy.xw} = 0$.

(4) $y$ is a linear expression in $x$, $z$ and $y$, i.e.

$$y = \alpha + \beta_x x + \sum \beta_{wi} w_i.$$

(5) $z$ and $w$ are independent, i.e. $\varrho_{zw_i} = 0$.

*Arrow scheme:*




*Aim:* Estimation of $\beta_x$.

*Material:* For every subject $x$, $y$, and $z$ are observed.

*Solution:* We multiply both the left and right sides of the equation by $(z - \bar{z})$ and construct mathematical expectations:

$$E\,y\,(z-\bar{z}) = E\,\alpha\,(z-\bar{z}) + E\,\beta_x\,x\,(z-\bar{z}) + E\sum \beta_{w_i}\,w_i\,(z-\bar{z}). \qquad (83)$$

Since $\varrho_{zw_i} = 0$ we derive that $Ew_i\,(z-\bar{z}) = 0$.

Hence the expression (83) can be simplified.

$$E\,y(z-\bar{z}) = \beta_x E\,x(z-\bar{z});$$

i.e.
$$\beta_x = \frac{E\,y\,(z-\bar{z})}{E\,x\,(z-\bar{z})} = \frac{E\,r_{yz}\,\sigma_y}{E\,r_{xz}\,\sigma_x}.$$

A consistent estimator of $\beta_x$, here called $b_x^R$, can be obtained by estimating the numerator and denominator by means of the observations. We find:

$$b_x^R = \frac{r_{yz}}{r_{xz}}\,\frac{s_y}{s_x} = \frac{\sum y\,(z-\bar{z})}{\sum x\,(z-\bar{z})}. \qquad (84)$$

*Remark:* The linearity in the models I–III may be of the type (1), (2) or (3) in the section 2.61.

**Models IV and V**

*Assumptions:* The same as in model III, except assumption 3. Here we may allow that $\varrho_{zy.xw_i} \neq 0$.

*Arrow scheme:*

The following equality is valid:

$$y = \alpha + \beta_x x + \beta_z z + \sum \beta_{w_i} w_i + \varepsilon,$$

where $E(\varepsilon) = 0$. We can in these models allow $\sigma(\varepsilon)$ to depend on $x$, $z$ and/or $w$. We can for convenience define $w_i$ so that $\beta_{w_i} \geqslant 0$.
The linearity must hold for individuals as in 2.61: (1).
In model IV it is further assumed that $z = 1$ is a necessary but not sufficient condition for $x = 1$. This assumption is modified in model V. We assume there only that $r_{zx} > 0$.

*Aim:* Estimation of $\beta_x$.

**Model IV**

*Material:* For every subject, values of $x$, $y$ and $z$ are known. The values for $x$ and $z$ can be combined in four different ways. We know how many subjects are included in every group, and we also know that the group where $x = 1$, and $z = 0$ is empty. For the remaining three groups the mean of $y$ can be calculated (cf. table).

TABLE

| Group | $x$ | $z$ | Number of subjects | Relative number of subjects where $z=0$ and $z=1$ respectively | Mean of $y$ |
|---|---|---|---|---|---|
| $G_0$ | 0 | 0 | $n_0$ | $p_0=\frac{n_0}{n_0+n_x}=1$ | $\bar{y}_0$ |
| $G_x$ | 1 | 0 | $n_x=0$ | $p_x=1-p_0=0$ | ? |
| $G_z$ | 0 | 1 | $n_z$ | $p_z=\frac{n_z}{n_z+n_{xz}}$ | $\bar{y}_z$ |
| $G_{zx}$ | 1 | 1 | $n_{zx}$ | $p_{zx}=1-p_z$ | $\bar{y}_{zx}$ |

*Approach for model IV (ignoring $\varepsilon$):*

The material is grouped according to the values of $x$ and $z$. If $z=0$ it is always the case that $x=0$ and the corresponding individuals belong to $G_0$. Individuals for which $x=0$ and $z=1$ belong to the group $G_z$, and individuals for which $x=1$ and $z=1$ belong to $G_{zx}$.

It is assumed that $z$ has no effect on $w_i$ or vice versa ($z$ and $w$ are independent). From this it follows (cf. 2.40), that the probability distribution of $w_i$ is the same for the sub-populations $z=0$ and $z=1$ (we ignore the $x$-values). The assumption can be reformulated:

$$\phi(w_0, w_1, \ldots, w_k|z=0)=\phi(w_0, w_1, \ldots, w_k|z=1), \tag{85}$$

where $\phi$ stands for a cumulative joint probability distribution function.

This assumption implies that to every individual belonging to the group $z=0$ and for which $w_0=1$, $w_1=0$, $w_2=1$ etc. there correspond $(n_z+n_{zx})/n_0$ individuals in the group $z=1$, having the same $w_0$, $w_1$ and $w_2$-values etc., i.e. as if matching were performed with respect to the $w_i$-values. Since we have no information concerning $w_i$, we do not know from the observation data which individuals in $G_z$ and $G_0$ correspond to each other; nor do we know which individuals in $G_{zx}$ and $G_0$ correspond. Hence we cannot find an exact estimator of $\beta_x$; we can, however, find max. and min. values for $\beta_x$.

In order to find a min. value let us assume that the $n_0 p_{zx}$ individuals in $G_0$ having the largest $y$-values correspond to the individuals in $G_{zx}$, i.e. the joint probability distributions with respect to the $w_i$-values are the same. Analogically, we assume that those having the smallest $y$-values correspond to the individuals in the $G_z$ group.

*Max. and min. values of* $\beta_z$: The min. and max. solutions are illustrated in Fig. 20 and Fig. 21, respectively. The solutions are possible in spite of the fact that $z$ has an effect on $y$.

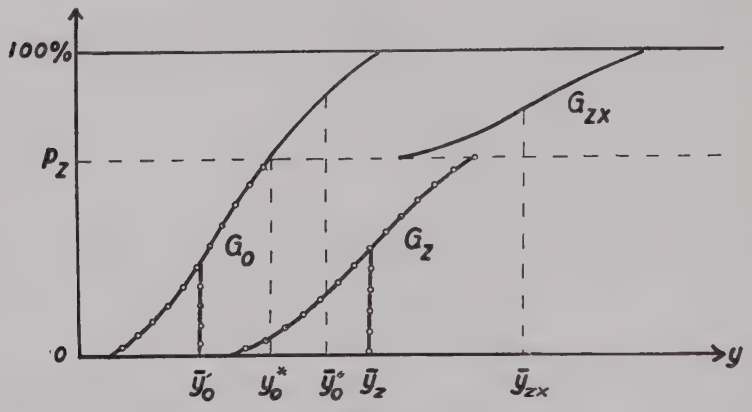



Fig. 20.

*Legend to Fig. 20:* The curve $G_z$ starts on the horizontal axis and is non-decreasing until it reaches the $p_z$-line. The curve $G_{zx}$ starts from a point on the $p_z$-line and reaches the 100 %-line. We assume that those individuals having $y$-values belonging to the part of $G_0$ lying below the $p_z$-line have the same joint probability distribution of $w_i$-values as the individuals in the $G_z$ group. In the same way the individuals within the $G_{zx}$ group correspond to those individuals having values belonging to the part of $G_0$ lying above the $p_z$-line. Fig. 20 represents the situation where the $w$-values have a strong positive effect on $y$ as well as on $x$.

We introduce the notation $\bar{y}_0'$ for the mean of the sub-group $G_0$ for which $y \leqslant y_0^*$, and $\bar{y}_0''$ for the sub-group $G_0$ for which $y > y^*$. By $\bar{y}_z$ and $\bar{y}_{zx}$ we denote the mean of the $y$-values for $G_z$ and $G_{zx}$ respectively.

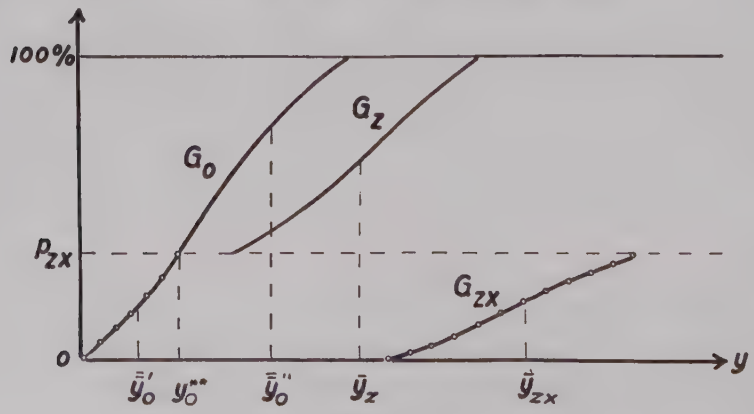



Fig. 21.

*Legend to Fig. 21:* The curve $G_{zx}$ starts on the horizontal axis and reaches the $p_{zx}$-line. $G_z$ starts on the $p_{zx}$-line and reaches the 100 %-line. Fig. 21 illustrates the situation where those combinations of $w$-values having the strongest effect on $y$ have the weakest effect on $x$. By $\bar{\bar{y}}_0'$ and $\bar{\bar{y}}_0''$ we denote the means of the $y$-values for those parts of $G_0$ which lie respectively below and above the $p_{zx}$-line.

From Fig. 20 we see that $\bar{y}_z - \bar{y}_0'$ is an estimate of the max. values for $\beta_z$; i.e. $E\bar{y}_z - E\bar{y}_0' \geqslant \beta_z$.

We further obtain:

$$\beta_x + \beta_z \geqslant E\bar{y}_{zx} - E\bar{y}_0''.$$

From these two inequalities we find

$$\beta_x \geqslant E(\bar{y}_{zx} - \bar{y}_0'' - \bar{y}_z + \bar{y}_0'); \tag{86}$$

i.e. a min. estimate for $\beta_x$ that can be calculated from the four mean values.

Similarly from Fig. 21 we find a max. value for $\beta_x$, since

$$\beta_x \leqslant E(\bar{y}_{zx} - \bar{\bar{y}}_0' - \bar{y}_z + \bar{\bar{y}}_0'') \tag{87}$$

*Remark:* We have assumed that no individuals belong to $G_x$. If instead we had assumed that either $G_0$, $G_z$ or $G_{zx}$ was empty, an analogous approach could have been used.

**Model V**

*Assumptions:* The same as in model IV. Let us choose $x = 1$ so that the relative frequency for $G_{zx}$ is larger than that for $G_x$; i.e. $p_{zx} > p_x$. Moreover we assume that

$$P(x = 1 \,|\, w, z = 0) \leqslant P(x = 1 \,|\, w, z = 1)$$

for every combination of $w$-values (cf. 2.66). The following relation is assumed to hold for every individual:

$$y = \beta_x x + \beta_z z + W(w) + \varepsilon.$$

*Observation material:* For every individual we know the $x$-, $y$- and $z$-values. The material can be grouped as in the table below.

| Group | $x$ | $z$ | Relative frequency | Mean of $y$ | Number of individuals |
|---|---|---|---|---|---|
| $G_0$ | 0 | 0 | $p_0$ | $\bar{Y}_0$ | $p_0 N_0 = n_0$ |
| $G_x$ | 1 | 0 | $p_x = 1 - p_0$ | $\bar{Y}_x$ | $p_x N_0 = n_x$ |
| $G_z$ | 0 | 1 | $p_z$ | $\bar{Y}_z$ | $p_z N_z = n_z$ |
| $G_{zx}$ | 1 | 1 | $p_{zx} = 1 - p_z$ | $\bar{Y}_{zx}$ | $p_{zx} N_z = n_{zx}$ |

*Max. and min. estimates for $\beta_x$:* The solution is demonstrated by help of the min. diagram in Fig. 22.

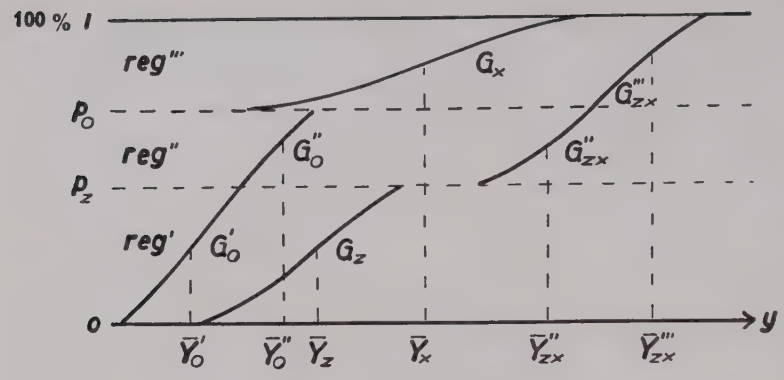



Fig. 22.

*Legend to Fig. 22:* The group $G_0$ is divided into two sub-groups $G_0'$ and $G_0''$. $G_{zx}$ is divided into $G_{zx}''$ and $G_{zx}'''$. Observe that $G_x$ and $G_{zx}$ reach the 100 %-line. The material is ordered in three regions. The means are denoted by a bar. By $\bar{Y}_{zx}'''$, for instance, is meant the mean for $G_{zx}'''$, i.e. the mean for those $(1-p_0)N_z$ individuals belonging to $G_{zx}$ who have the largest $y$-values.

*Notations:*

(i) We introduce $\mu_{zx}'''$ to denote the mean $y$-value of those $(1-p_0)N_z$ individuals belonging to $G_{zx}$ who have the largest $(y-\varepsilon)$-values, i.e. the largest $y$-values after the effect of the residual factor $\varepsilon$ is ignored (we can call $\mu_{zx}'''$ the true mean value). $\mu_{zx}''$ corresponds to the lowest $(y-\varepsilon)$-values in $G_{zx}$ etc.

(ii) The relation between $x$ and $w$ is denoted in the following way. Given $z=0$, the probability for $x=1$ can be expressed as a function of $w$: $P(x=1\,|\,w, z=0)=g(w)$. The probability for $x=0$ is then $P(x=0\,|\,w, z=0)=1-g(w)$. The corresponding probabilities in the case $z=1$ are denoted $P(x=1\,|\,w, z=1)=h(w)$ and $P(x=0\,|\,w, z=1)=1-h(w)$ respectively.

As they are probabilities, the inequalities

$$0 \leqslant g(w) \leqslant 1 \text{ and } 0 \leqslant h(w) \leqslant 1$$

hold for every value of $w$.

(iii) The symbols $p_0$, $p_x$, $p_z$, and $p_{zx}$ have the same meaning as in the table. Now we can derive (the integration is taken over all $w$-values):

$$\int g(w)\, d\phi_w(w) = p_x; \quad \int \{1-g(w)\}\, d\phi_w(w) = p_0 = 1-p_x. \tag{88}$$

$$\int h(w)\, d\phi_w(w) = p_{zx}; \quad \int \{1-h(w)\}\, d\phi_w(w) = p_z = 1-p_{zx}. \tag{89}$$

*Derivation of a min. value for* $\beta_x$: Since $E(\varepsilon) = 0$ we obtain:

$$E\overline{Y}_x = \mu_x,\ E\overline{Y}_{zx} = \mu_{zx},\ E\overline{Y}_0 = \mu_0,\ E\overline{Y}_z = \mu_z. \tag{90}$$

Moreover the following inequalities must be valid (cf. p. 127):

$$E\overline{Y}'''_{zx} \geqslant \mu'''_{zx},\ E\overline{Y}''_{zx} \leqslant \mu''_{zx},\ E\overline{Y}''_0 \geqslant \mu''_0,\ E\overline{Y}'_0 \leqslant \mu'_0. \tag{91}$$

We can derive a relation between $\mu'''_{zx}$ and $\mu_x$. We have assumed that the inequality $P(x = 1 \mid w, z = 0) \leqslant P(x = 1 \mid w, z = 1)$ holds for every $w$, i.e.

$$g(w) \leqslant h(w). \tag{92}$$

Further we know that $E\overline{Y}_x$ equals (approximately):

$$E(y|x = 1, z = 0) = \beta_x + \int W(w)\, g(w)\, d\phi_w(w)/p_x = \mu_x. \tag{93}$$

From (92) it follows that we can choose $p_x$ individuals in $G_{zx}$ for which the true expected $y$-mean is (at least) $\beta_z + \beta_x + \int W(w) g(w) d\phi_w(w)/p_x$. If all these individuals have larger $y$-values than the remaining individuals in $G_{zx}$ this expression can be equal to $\mu'''_{zx}$, in other cases $\mu''''_{zx}$ is larger.

We now obtain by means of eq. (93) the following relation between $\mu''''_{zx}$ and $\mu_x$.

$$E(\overline{Y}_x) + \beta_z = \mu_x + \beta_z = \beta_x + \beta_z + \int W(w)\, g(w)\, d\phi_w(w)/p_x \leqslant \mu''''_{zx}; \tag{94}$$

i.e.

$$\mu_x + \beta_z \leqslant \mu''''_{zx}.$$

From the inequalities (91) it follows that

$$E\overline{Y}'''_{zx} - E\overline{Y}_x \geqslant \mu''''_{zx} - \mu_x \geqslant \beta_z. \tag{95}$$

From analogous arguments we obtain

$$\mu''_{zx} - \beta_z - \beta_x \leqslant \mu''_0. \tag{96}$$

From eqs. (91) and (96) it then follows

$$\beta_z + \beta_x \geqslant \mu''_{zx} - \mu''_0 \geqslant E\overline{Y}''_{zx} - E\overline{Y}''_0. \tag{97}$$

From the relations (95) and (97) we obtain the following min. estimate for $\beta_x$:

$$\overline{Y}''_{zx} - \overline{Y}''_0 - \overline{Y}'''_{zx} + \overline{Y}_x.$$

**Model VI**

*Arrow scheme:*

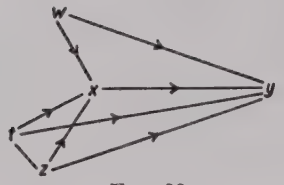



Fig. 23.

*Assumptions:*

(1) We assume that values for two instrumental variables $z$ and $t$ are known, and that only the following combinations of values are possible:

$$\begin{cases} z=0 \\ t=0 \end{cases} \qquad \begin{cases} z=1 \\ t=0 \end{cases} \qquad \begin{cases} z=1 \\ t=1 \end{cases}$$

These three combinations of values can correspond to three different levels of a factor.

(2) We further assume that no interactions including $x$ exist, i.e. $\gamma_{zx}=0$. $\gamma_{tx}=0$, $\gamma_{wix}=0$ for all $w_i$ etc. (cf. 2.63).

The hypothetical relation between $y$ and the explaining variables can be written:

$$y=\alpha+\beta_x x+\beta_z z+\beta_t t+\sum\beta_{w_i}w_i+\sum\gamma_{zw_i}zw_i+\sum\gamma_{tw_i}tw_i+\sum\gamma_{w_iw_j}w_iw_j+$$
$$+\text{ higher orders of interaction, except those in which } x \text{ is included.}$$

$\varepsilon$ is not included in this relation. We introduce the symbol $Y$ for denoting observed values for the dependent variable where $Y=y+\varepsilon$.

(3) We assume that $z$ as well as $t$ is independent of each $w_i$ (cf. 2.41).

(4) Further we assume that the interactions between $z$ and $w_1, \ldots, w_k$ and those between $t$ and $w_1, \ldots, w_k$ are related. We specify this assumption in detail on p. 121.

*Observation material:* For every individual, $x, Y, z$, and $t$ are observed. The observation material can be grouped as in the table.

| Group | $z$ | $t$ | $x$ | Number of individuals | Relative frequency | Observed mean |
|---|---|---|---|---|---|---|
| (0) | 0 | 0 | 0 | $N_0$ | $\frac{N_0}{N_0}=1$ | $\overline{Y}_0$ |
| ($z$) | 1 | 0 | 0 | $N_z$ | $p_z=\frac{N_z}{N_z+N_{zx}}$ | $\overline{Y}_z$ |
| ($zx$) | 1 | 0 | 1 | $N_{zx}$ | $p_{zx}=1-p_z$ | $\overline{Y}_{zx}$ |
| ($ztx$) | 1 | 1 | 1 | $N_{ztx}$ | $\frac{N_{ztx}}{N_{ztx}}=1$ | $\overline{Y}_{ztx}$ |

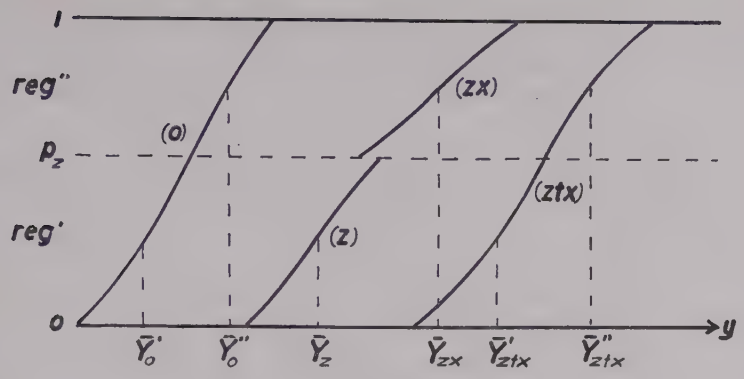



FIG. 24.

Note that in group (0) $x$ is always equal to zero and in group $(ztx)$ $x$ is always equal to one.

*Cumulative frequency diagram:* The diagram illustrates an extreme situation. The curve $(z)$ has been placed below the $p_z$-line and the curve $(zx)$ above this line. The symbols $\bar{Y}_0'$, $\bar{Y}_0''$, $\bar{Y}_z'$ etc. refer to the observed means of the $Y$-values belonging to the respective curves and regions.

*Solution:* We begin by attempting to prove that it is possible to find a min. estimator for $\beta_x$ under the given assumptions. The proof is carried out in three stages.

(1) We specify some assumptions concerning the interactions between the instrumental variables and the $w$-factors.

(2) We ignore the $\varepsilon$-factor and construct a min. diagram corresponding to the most extreme correlations between $x$, $y$ and $w$ (see Fig. 24).

(3) We show that it is possible to find min. values even if a residual factor $\varepsilon$ which is independent of $w$, $z$, $t$, and $x$ is not ignored. The demonstration is valid even if the dispersion for $\varepsilon$ depends on $y$, $w$, $z$, $t$, or $x$.

*Stage 1* (Assumption (4) on p. 118 is specified):

Ignoring the fact that $w$ has an effect on $x$ and that $\varepsilon$ exists we would find that if $z = 1$, $t = 0$, and $x = 0$:

$$y = \alpha + \beta_z + \sum \beta_{wi} w_i + \sum \gamma_{zwi} z w_i + \sum \gamma_{wiwj} w_i w_j \text{ etc. } (\varepsilon \text{ not included}). \quad (98)$$

We introduce $w$ and illustrate how $y$ can be described as a function of a rectangularly distributed variable $w$ by a cumulative frequency diagram (see Fig. 25).

We denote the vertical axis as the $w$-axis, and we can choose $w$ so that $0 \leqslant w \leqslant 1$. The curve denoted as $y = G_z(w)$ is non-decreasing. The following relations between $y = G_z(w)$ and (98) is introduced for defining $w$:

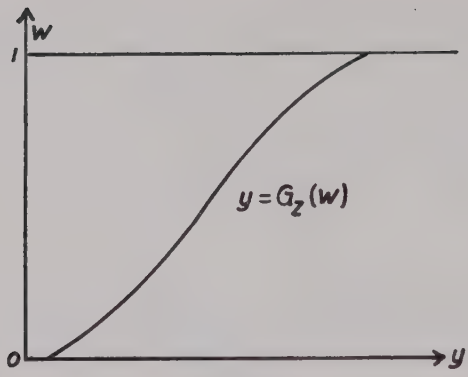



FIG. 25.

$$w = G_z^{-1}(y) = \int_{y \leqslant \alpha + \beta_z + \Sigma \gamma_{zw_i} w_i + \Sigma \gamma_{w_i w_j} w_i w_j \text{ etc.}} \varphi(w_1, w_2, \ldots, w_k)\, d w_1\, d w_2 \ldots d w_k. \tag{99}$$

($\varphi$ denotes the joint frequency function for $w_1, w_2, \ldots, w_k$).

By this transformation $w$ becomes a rectangularly distributed variable in the region (0, 1). To a point on $G_z(w)$ just above $w = 0$ there corresponds the combination of values, $w_1, w_2, \ldots, w_k$, giving the smallest possible $y$-value and to a point just below $w = 1$, the combination giving the greatest $y$-value. The combinations are ordered with respect to their effect on $y$.

We introduce the notations $G_0$ and $G_{zt}$ in the table; $w$ is defined in the relation (99).

| $z$ | $t$ | $x$ | $y$ expressed as a function of $w$ and $x$ ($\varepsilon$ is ignored) |
|---|---|---|---|
| 0 | 0 | 0 | $y = G_0(w)$ |
| 1 | 0 | 0 or 1 | $y = G_z(w) + \beta_x x$ |
| 1 | 1 | 1 | $y = G_{zt}(w) + \beta_x$ |

We introduce $\Delta_z(w)$ and $\Delta_t(w)$ denoting the following differences:

$$\begin{aligned} G_z(w) - G_0(w) &= \Delta_z(w), \\ G_{zt}(w) - G_z(w) &= \Delta_t(w). \end{aligned} \tag{100}$$

*Assumptions and notations:*

(i) $G_0$, $G_z$, and $G_{zt}$ are finite and can be differentiated.

(ii) $G_z(w)$ is a non-decreasing function, i.e. $G_z'(w) \geqslant 0$. (This assumption need not hold for $G_0'$ and $G_{zt}'$.)

(iii) $w$ is a rectangularly distributed variable in the interval (0, 1).

(iv) (Cf. assumption 4 on p. 118). The signs (positive or negative) for the following expressions are assumed to be the same:

$$\text{Sign}\left\{\int_0^p \Delta_z(w)\,dw - p\int_0^1 \Delta_z(w)\,dw\right\} = \text{Sign}\left\{\int_0^p \Delta_t(w)\,dw - p\int_0^1 \Delta_t(w)\,dw\right\} \quad (101)$$

holds for all $p$, where $0 \leqslant p \leqslant 1$.

*Remark 1:* We introduce $\bar{\Delta}_z$ and $\bar{\Delta}_t$ defined:

$$\int_0^1 \Delta_z(w)\,dw = \bar{\Delta}_z \quad \text{and} \quad \int_0^1 \Delta_t(w)\,dw = \bar{\Delta}_t.$$

*Remark 2:* The difference (cf. 101)

$$\int_0^p \Delta_z(w)\,dw - p\int_0^1 \Delta_z(w)\,dw = \int_0^p (\Delta_z(w) - \bar{\Delta}_z)\,dw$$

is the interaction between $w$ and $z$ expressed as a function of $p$. If $p = 1$ this expression is chosen so that it equals zero; i.e. we standardize the interaction. In the same way we construct the expression

$$\int_0^p (\Delta_t(w) - \bar{\Delta}_t)\,dw.$$

*Remark 3:* The assumption concerning the sign (101) implies that the two interactions are related. This assumption is sometimes plausible if $z$ and $t$ are related.

*Remark 4:* Assumption (iv) can also be written:

$$\text{Sign}\int_p^1 (\Delta_z(w) - \bar{\Delta}_z)\,dw = \text{Sign}\int_p^1 (\Delta_t(w) - \bar{\Delta}_t)\,dw.$$

This is obvious from remark 2.

*Stage 2* (Comment to Fig. 26):

$G_0(w)$, $G_z(w)$, $G_z(w) + \beta_x$ and $G_{zt}(w) + \beta_x$ are illustrated in the figure. Note that $G_z(w)$ and $G_z(w) + \beta_x$ are non-decreasing.

In the figure the lowest $y$-values on the respective curves have been marked by *xxxx*. We choose these *xxxx* parts of the $G_z$ curve so that if projected on the $w$-axis they correspond to an interval of the length $p_z$ (0.6 in the figure). The corresponding length of the projection of all *xxxx* from $G_{zt}$ and $G_0$ is also $p_z$.

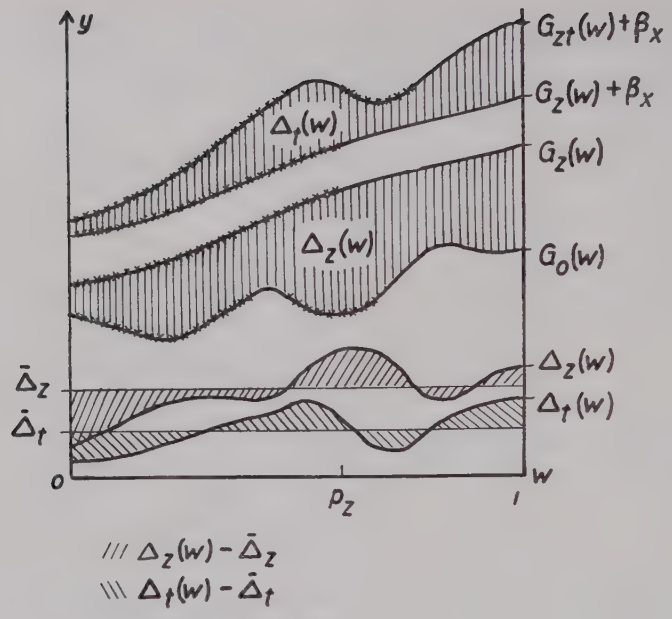



FIG. 26.

We denote the (expected) mean of $y$ and introduce the notation $\mu$ in the following way (note that $\varepsilon$ is ignored):

| Curve | The *xxxx* part of the curve | The other part of the curve | Mean in region $0 \leqslant w \leqslant p_z$ | Mean in region $p_z \leqslant w \leqslant 1$ |
|---|---|---|---|---|
| $G_{zt}(w)+\beta_x$ | $\bar{y}'_{zt}+\beta_x$ | $\bar{y}''_{zt}+\beta_x$ | $\mu'_{zt}+\beta_x$ | $\mu''_{zt}+\beta_x$ |
| $G_z(w)+\beta_x$ | $\bar{y}'_z+\beta_x$ | $\bar{y}''_z+\beta_x$ | $\mu'_z+\beta_x$ | $\mu''_z+\beta_x$ |
| $G_z(w)$ | $\bar{y}'_z$ | $\bar{y}''_z$ | $\mu'_z$ | $\mu''_z$ |
| $G_0(w)$ | $\bar{y}'_0$ | $\bar{y}''_0$ | $\mu'_0$ | $\mu''_0$ |

We introduce the further notations $A$ and $B$:

$$A \equiv \bar{y}''_z - \bar{y}'_z + \bar{y}'_0 - \bar{y}''_0; \quad B \equiv \bar{y}'_{zt} - \bar{y}''_{zt} + \bar{y}''_z - \bar{y}'_z.$$

*Aim:* We want to prove that

if $$A \equiv \bar{y}''_z - \bar{y}'_z + \bar{y}'_0 - \bar{y}''_0 \geqslant 0$$

then $$B \equiv \bar{y}'_{zt} - \bar{y}''_{zt} - \beta_x + \bar{y}''_z + \beta_x - \bar{y}'_z \leqslant 0.$$

*Proof:* We know that (cf. eq. 100):

$$G_z(w) = G_0(w) + \Delta_z(w)$$

and that

$$G_{zt}(w) = G_z(w) + \Delta_t(w).$$

Since $G_z(w)$ is a non-decreasing function the mean of the *xxxx* marked part of $G_z(w)$ can be expressed:

$$\frac{\int_0^{p_z} G_z(w)\, dw}{p_z} = \bar{y}_z' = \mu_z',$$

while the mean of the region not marked with crosses is

$$\frac{\int_{p_z}^{1} G_z(w)\, dw}{1-p_z} = \bar{y}_z'' = \mu_z''.$$

We can now write $A$:

$$A = \bar{y}_z'' - \bar{y}_z' + \bar{y}_0' - \bar{y}_0'' = \frac{\int_{p_z}^{1} (G_0(w) + \Delta_z(w))\, dw}{1-p_z} - \frac{\int_0^{p_z} (G_0(w) + \Delta_z(w))\, dw}{p_z} + \bar{y}_0' - \bar{y}_0''. \tag{102}$$

Since it is not certain that all the lower $y$-values lie in the interval $0 \leqslant w \leqslant p_z$, it is obvious that

$$\bar{y}_0' \leqslant \frac{\int_0^{p_z} G_0(w)\, dw}{p_z} = \mu_0'. \tag{103}$$

In the same way we obtain:

$$\bar{y}_0'' \geqslant \frac{\int_{p_z}^{1} G_0(w)\, dw}{1-p_z} = \mu_0''. \tag{104}$$

From the relations (103) and (104) it follows:

$$\bar{y}_0' - \bar{y}_0'' \leqslant \frac{\int_0^{p_z} G_0(w)\, dw}{p_z} - \frac{\int_{p_z}^{1} G_0(w)}{1-p_z}. \tag{105}$$

If this inequality is applied to (102) we obtain

$$A \leqslant \frac{\int_{p_z}^{1} (G_0(w) + \Delta_z(w))\, dw}{1-p_z} - \frac{\int_{0}^{p_z} (G_0(w) + \Delta_z(w)\, dw}{p_z} + \\ + \frac{\int_{0}^{p_z} G_0(w)\, dw}{p_z} - \frac{\int_{p_z}^{1} G_0(w)\, dw}{1-p_z}. \tag{106}$$

Simplifying (106) we obtain

$$A \leqslant \frac{\int_{p_z}^{1} \Delta_z(w)\, dw}{1-p_z} - \frac{\int_{0}^{p_z} \Delta_z(w)\, dw}{p_z}, \tag{107}$$

which can be written (cf. remark 2, p. 121):

$$A \leqslant \frac{\int_{p_z}^{1} (\Delta_z(w) - \bar{\Delta}_z) dw}{1-p_z} + \bar{\Delta}_z - \frac{\int_{0}^{p_z} (\Delta_z(w) - \bar{\Delta}_z)\, dw}{p_z} - \bar{\Delta}_z \tag{108}$$

$$= \frac{\int_{p_z}^{1} (\Delta_z(w) - \bar{\Delta}_z) dw}{1-p_z} - \frac{\int_{0}^{p_z} (\Delta_z(w) - \bar{\Delta}_z)\, dw}{p_z}. \tag{109}$$

We have defined $\bar{\Delta}_z$ so that $\int_0^1 (\Delta_z(w) - \bar{\Delta}_z)\, dw = 0$ (cf. remark 1, p. 121). From this it follows that $\int_0^{p_z} (\Delta_z - \bar{\Delta}_z)\, dw = -\int_{p_z}^1 (\Delta_z - \bar{\Delta}_z)\, dw$. If this expression is substituted in (109) we obtain:

$$A \leqslant \int_{p_z}^{1} (\Delta_z(w) - \bar{\Delta}_z) dw \left( \frac{1}{1-p_z} + \frac{1}{p_z} \right). \tag{110}$$

If $A \geqslant 0$ the following inequality holds: $0 \leqslant \int_{p_z}^{1} (\Delta_z(w) - \bar{\Delta}_z)\, dw$.

We proceed by simplifying $B \equiv \bar{y}'_{zt} - \bar{y}''_{zt} + \bar{y}''_z - \bar{y}'_z$ given that $A \geqslant 0$. Hence it follows from (110) that $\int_{p_z}^{1} (\Delta_z(w) - \bar{\Delta}_z)\, dw \geqslant 0$. Analogically to (103) and (104) we obtain

$$\bar{y}'_{zt} \leqslant \frac{\int_0^{p_z} (G_z(w) + \Delta_t(w))\, dw}{p_z} = \frac{\int_0^{p_z} G_{zt}(w)\, dw}{p_z} = \mu'_{zt}, \tag{111}$$

$$\bar{y}''_{zt} \geqslant \frac{\int_{p_z}^{1} (G_z(w) + \Delta_t(w))\, dw}{1 - p_z} = \frac{\int_{p_z}^{1} G_{zt}(w)\, dw}{1 - p_z} = \mu''_{zt}. \tag{112}$$

From (111) and (112) it follows that

$$B = \bar{y}'_{zt} - \bar{y}''_{zt} + \bar{y}''_z - \bar{y}'_z \leqslant \frac{\int_0^{p_z} (G_z(w) + \Delta_t(w))\, dw}{p_z} - \frac{\int_{p_z}^{1} (G_z(w) + \Delta_t(w))\, dw}{1 - p_z} +$$

$$+ \frac{\int_{p_z}^{1} G_z(w)\, dw}{1 - p_z} - \frac{\int_0^{p_z} G_z(w)\, dw}{p_z} = \mu'_{zt} - \mu''_{zt} + \mu''_z - \mu'_z$$

$$= \frac{\int_0^{p_z} \Delta_t(w)\, dw}{p_z} - \frac{\int_{p_z}^{1} \Delta_t(w)\, dw}{1 - p_z} = \frac{\int_0^{p_z} (\Delta_t(w) - \bar{\Delta}_t)\, dw}{p_z} - \frac{\int_{p_z}^{1} (\Delta_t(w) - \bar{\Delta}_t)\, dw}{1 - p_z}$$

$$= -\int_{p_z}^{1} (\Delta_t(w) - \bar{\Delta}_t)\, dw \left( \frac{1}{1 - p_z} + \frac{1}{p_z} \right) \leqslant 0.$$

The last inequality follows from assumption (iv) p. 121 and from the fact that $A \geqslant 0$; i.e. if $A \geqslant 0$, it follows that $B \leqslant 0$.

*Conclusion:* Either $A \leqslant 0$ or $B \leqslant 0$.

*Estimation of min. value for $\beta_x$ if $\varepsilon$ is ignored.*

*Proof:* The proof demonstrates how a min. value for $\beta_x$ is obtained if $\varepsilon$ is ignored. According to the above proof we know:

either $$A = \bar{y}''_z - \bar{y}'_z + \bar{y}'_0 - \bar{y}''_0 \leqslant 0, \tag{113}$$

or $$B = \bar{y}'_{zt} - \bar{y}''_{zt} + \bar{y}''_z - \bar{y}'_z \leqslant 0. \tag{114}$$

We have mentioned earlier that $w \to -x$. When the instrumental variables have the values $z = 1$ and $t = 0$ the material can be divided into two groups $x = 0$ and $x = 1$, the relative frequency being $p_z$ and $1 - p_z$ respectively. We introduce the conditional probability $P(x = 1 \,|\, w, z = 1, t = 0) = h(w)$; $h(w)$ is introduced since the probability is a function of $w$. The following relation holds

$$\int_0^1 P(x=1 \mid w,\ z=1,\ t=0)\, dw = \int_0^1 h(w)\, dw = p_{zx} = 1 - p_z.$$

We add the *notations*: $\bar{y}_{zx}$ = expected mean of $y$ for $x = 1$ and $z = 1$ if $\varepsilon$ is ignored and $\bar{y}_z$ = expected mean of y for $x = 0$ and $z = 1$ if $\varepsilon$ is ignored.
We assume that these two means have been obtained by the help of the curves $G_z(w) + \beta_x$ and $G_z(w)$ respectively.
For $\bar{y}_{zx}$ and $\bar{y}_z$ it holds:

$$\bar{y}_{zx} = \frac{\int_0^1 h(w)\,(G_z(w) + \beta_x)\, dw}{1 - p_z} = \beta_x + \frac{\int_0^1 h(w)\, G_z(w)\, dw}{1 - p_z};$$

$$\bar{y}_z = \frac{\int_0^1 (1 - h(w))\, G_z(w)\, dw}{p_z}.$$

We construct

$$\bar{y}_{zx} - \bar{y}_z - \beta_x - \bar{y}_z'' + \bar{y}_z'$$

$$= \frac{\int_0^1 h(w)\, G_z(w)\, dw}{1 - p_z} - \frac{\int_0^1 (1 - h(w))\, G_z(w)\, dw}{p_z} - \frac{\int_{p_z}^1 G_z(w)\, dw}{1 - p_z} + \frac{\int_0^{p_z} G_z(w)\, dw}{p_z}$$

$$= \frac{\int_0^1 G_z(w)\, dw}{1 - p_z} - \int_0^1 (1 - h(w))\, G_z(w)\, dw \left(\frac{1}{1 - p_z} + \frac{1}{p_z}\right) -$$

$$- \frac{\int_{p_z}^1 G_z(w)\, dw}{1 - p_z} + \frac{\int_0^{p_z} G_z(w)\, dw}{p_z}$$

$$= \int_0^{p_z} G_z(w)\, dw \left(\frac{1}{1 - p_z} + \frac{1}{p_z}\right) - \int_0^1 (1 - h(w))\, G_z(w) \left(\frac{1}{1 - p_z} + \frac{1}{p_z}\right)$$

$$= \left(\frac{1}{1 - p_z} + \frac{1}{p_z}\right) \left\{ \int_0^{p_z} G_z(w)\, dw - \int_0^1 (1 - h(w))\, G_z(w)\, dw \right\}.$$

This relation equals zero if and only if $h(w) = 0$ when $0 \leqslant w \leqslant p_z$ and $h(w) = 1$ when $p_z \leqslant w \leqslant 1$.

In other cases:

$$\bar{y}_{zx} - \bar{y}_z - \beta_x - \bar{y}_z'' + \bar{y}_z' \leqslant 0.$$

By substituting $\bar{y}_z'' - \bar{y}_z'$ in (113) and (114) the following conclusion is reached:

*Conclusion:* If $A < 0$ it follows that

$$\bar{y}_{zx} - \bar{y}_z - \beta_x + \bar{y}_0' - \bar{y}_0'' \leqslant A \leqslant 0;$$

and if $B < 0$ it follows that

$$\bar{y}_{zx} - \bar{y}_z - \beta_x + \bar{y}_{zt}' - \bar{y}_{zt}'' \leqslant B \leqslant 0.$$

From these relations we can derive two min. values for $\beta_x$. Since either $A \leqslant 0$ or $B \leqslant 0$ we should choose the lesser of these min. values.

*Stage 3:*

We now proceed to demonstrate what happens if the residual factor $\varepsilon$ is not ignored. Let us add $\varepsilon$ to $G_z(w)$. We denote the observed values of the dependent variable by $Y$ (cf. Fig. 24), where $Y = y + \varepsilon$. By $y$ we mean the $y$-value before $\varepsilon$ is taken into account. Note that $E(\varepsilon) = 0$. Since $w$ and $\varepsilon$ are independent it follows that $E(\varepsilon \mid w) = 0$ for every $w$. Let us choose a fraction $p_z$ of the $G_z$-curve consisting of the lowest $y$-values and construct the mean, $\bar{y}_z'$; i.e. we choose

$$\bar{y}_z' = \frac{\int_0^{p_z} G_z(w)\, dw}{p_z} \quad \text{(cf. Fig. 26).}$$

If we had chosen a fraction consisting of the $p_z$ lowest $Y$-values we would have found that the mean of these values, $E\,\overline{Y}_z'$, is smaller than $\bar{y}_z'$ due to the $\varepsilon$-variation, i.e. $E\,\overline{Y}_z' \leqslant \bar{y}_z'$.

By an analogous argument we can prove that

$E\,\overline{Y}_0' \leqslant \bar{y}_0'$, where the means are based on fractions of the same size and $E\,\overline{Y}_{ztx}' \leqslant \bar{y}_{ztx}'$, where the means are based on fractions of the same size. $\overline{Y}_0'$ and $Y_{ztx}'$ correspond to the lowest $Y$-values in the groups (0) and $(ztx)$ in Fig. 24 respectively, and $\bar{y}_0'$ and $\bar{y}_{ztx}'$ correspond to the lowest values on the curves $G_0(w)$ and $G_{ztx}(w)$ in Fig. 26 respectively.

Since $E(p_0\,\overline{Y}_0' + (1-p_0)\,\overline{Y}_0'') = p_0\,\bar{y}_0' + (1-p_0)\,\bar{y}_0''$, if $E(\varepsilon) = 0$, it holds that

$$E\,\overline{Y}_0'' \geqslant \bar{y}_0'' \quad \text{(since } E\,\overline{Y}_0' \leqslant \bar{y}_0'\text{)}. \tag{115}$$

In the same way it holds that

$$E\,\overline{Y}_{zt}'' \geqslant \bar{y}_{zt}''.$$

Hence

$$E\,\overline{Y}_0'' - E\,\overline{Y}_0' \geqslant \bar{y}_0'' - \bar{y}_0', \tag{116}$$

and

$$E\,\overline{Y}_{ztx}'' - E\,\overline{Y}_{ztx}' \geqslant \bar{y}_{zt}'' - \bar{y}_{zt}'. \tag{117}$$

*Summary*

From the above inequalities we can obtain a min. value for $\beta_x$.

A) If $A \leqslant 0$, then

$$\bar{y}_{zx} - \beta_x - \bar{y}_z + \bar{y}_0' - \bar{y}_0'' \leqslant 0.$$

Since $E(\varepsilon) = 0$ we have

$$\bar{y}_{zx} = E\,\overline{Y}_{zx} \text{ and } \bar{y}_z = E\,\overline{Y}_z.$$

Further we have found (116)

$$E\,\overline{Y}_0'' - E\,\overline{Y}_0' \geqslant \bar{y}_0'' - \bar{y}_0',$$

i.e.

$$\beta_x \geqslant E\,(\overline{Y}_{zx} - \overline{Y}_z + \overline{Y}_0' - \overline{Y}_0'') \text{ if } A \leqslant 0. \tag{118}$$

B) If $A \geqslant 0$, and hence $B \leqslant 0$, we find in the same way that

$$E\,(\overline{Y}_{zx} - \beta_x - \overline{Y}_z + \overline{Y}_{ztx}' - \overline{Y}_{ztx}'') \leqslant 0$$

or

$$\beta_x \geqslant E\,(\overline{Y}_{zx} - \overline{Y}_z + \overline{Y}_{ztx}' - \overline{Y}_{ztx}'') \text{ if } A \geqslant 0. \tag{119}$$

We have now two possible estimators of the min. value, $\overline{Y}_{zx} - \overline{Y}_z + \overline{Y}_0' - \overline{Y}_0''$ and $\overline{Y}_{zx} - \overline{Y}_z + \overline{Y}_{ztx}' - \overline{Y}_{ztx}''$ respectively, and should choose the smaller one of these as the $\beta_x$ min. value.[1]

**Model VII**

*Assumptions:* We assume that the following relation holds for every individual (cf. 2.61) in a population:

$$y = \beta_x x + \beta_z z + W(w) + \varepsilon, \text{ where } E\,\varepsilon = 0. \tag{120}$$

Moreover we assume that $z$ and $w$ are independent, and the same is assumed for $z$ and $\varepsilon$, $x$ and $\varepsilon$, and $w$ and $\varepsilon$. Then the same arrow schemes as those in models IV and V are applicable. The variance of $\varepsilon$, $D^2(\varepsilon)$, is assumed to be independent of $x$, $z$, $w$ and $y$ (cf. 2.7). The function $W(w)$ is unknown.

*Observation material:* For every individual we know $x$, $y$ and $z$. We can divide the material as before into four groups (see the table on p. 129).

Every group is so large that sampling errors can be ignored. The solution is based on the special assumption that the residual variances are the same for each group.

[1] By analogous arguments a max. value solution is obtained as either

$$\overline{Y}_{zx} - \overline{Y}_z - \overline{Y}_0' + \overline{Y}_0'' \quad \text{or} \quad \overline{Y}_{zx} - \overline{Y}_z - \overline{Y}_{ztx}' + \overline{Y}_{ztx}''$$

| Group | $x$ | $z$ | Relative frequency | Observed means | Observed dispersions |
|---|---|---|---|---|---|
| $G_0$ | 0 | 0 | $p_0$ | $\overline{Y}_0$ | $s_0^2$ |
| $G_x$ | 1 | 0 | $p_x = 1 - p_0$ | $\overline{Y}_x$ | $s_x^2$ |
| $G_z$ | 0 | 1 | $p_z$ | $\overline{Y}_z$ | $s_z^2$ |
| $G_{zx}$ | 1 | 1 | $p_{zx} = 1 - p_z$ | $\overline{Y}_{zx}$ | $s_{zx}^2$ |

*Notations*: See models IV and V.

By $\overline{Y}_0$ we denote the mean and by $s_0^2$ the variance in the group $G_0$; by $\overline{Y}_x$ and $s_x^2$ corresponding values for $G_x$ etc. $g(w)$ and $h(w)$ are defined on p. 116. $\phi(w)$ is the cumulative distribution function for $w$.

*Solution:* We can now construct the mathematical expectations of $y$ for each of the four groups:

$$E(y|x=0,\ z=0) = \int \{1 - g(w)\}\, W(w)\, d\phi_w(w)/p_0 = E\overline{Y}_0, \tag{121}$$

$$E(y|x=1,\ z=0) = \beta_x + \int g(w)\, W(w)\, d\phi_w(w)/p_x = E\overline{Y}_x, \tag{122}$$

$$E(y|x=0,\ z=1) = \beta_z + \int \{1 - h(w)\}\, W(w)\, d\phi_w(w)/p_z = E\overline{Y}_z, \tag{123}$$

$$E(y|x=1,\ z=1) = \beta_x + \beta_z + \int h(w)\, W(w)\, d\phi_w(w)/p_{zx} = E\overline{Y}_{zx}, \tag{124}$$

where the integrals are taken over all $w$-values. Since $E(\varepsilon)$ equals 0, it does not appear in the above expressions. Since $w$ and $\varepsilon$ are independent, we can express the variances in the following way:

$$D^2(y|x=0,\ z=0) = \int W^2(w)\{1 - g(w)\}\, d\phi_w(w)/p_0 - \left\{\int \{1 - g(w)\}\, W(w)\, d\phi_w(w)/p_0\right\}^2 + D^2(\varepsilon) = E s_0^2, \tag{125}$$

$$D^2(y|x=1,\ z=0) = \int W^2(w)\, g(w)\, d\phi_w(w)/p_x - \left\{\int g(w)\, W(w)\, d\phi_w(w)/p_x\right\}^2 + D^2(\varepsilon) = E s_x^2, \tag{126}$$

$$D^2(y|x=0,\ z=1) = \int W^2(1-h)\, d\phi_w/p_z - \left\{\int (1-h)\, W\, d\phi_w/p_z\right\}^2 + D^2(\varepsilon) = E s_z^2, \tag{127}$$

$$D^2(y|x=1,\ z=1) = \int W^2 h\, d\phi_w/p_{zx} - \left\{\int h\, W\, d\phi_w/p_{zx}\right\}^2 + D^2(\varepsilon) = E s_{zx}^2. \tag{128}$$

We construct the expression:

$p_{zx}(128) + p_z(127) - p_x(126) - p_0(125)$, and simplify by substitution from eqs. (121)–(124):

$$-1/p_{zx}\left\{\int h\, W\, d\,\phi_w\right\}^2 - 1/p_z\left\{\int (1-h)\, W\, d\,\phi_w\right\}^2 + 1/p_x\left\{\int g\, W\, d\,\phi_w\right\}^2 +$$

$$+1/p_0\left\{\int (1-g)\, W\, d\,\phi_w\right\}^2 = -p_{zx}\left\{E\,(\overline{Y}_{zx}) - \beta_x - \beta_z\right\}^2 - p_z\left\{E\,(\overline{Y}_z) - \beta_z\right\}^2 +$$

$$+ p_x\left\{E\,(\overline{Y}_x) - \beta_x\right\}^2 + p_0\left\{E\,(\overline{Y}_0)\right\}^2.$$

This expression is equal to $p_{zx} E s_{zx}^2 + p_z E s_z^2 - p_x E s_x^2 - p_0 E s_0^2$.
Hence we find that

$$p_{zx} E^2 \overline{Y}_{zx} + p_z E^2 \overline{Y}_z - p_x E^2 \overline{Y}_x - p_0 E^2 \overline{Y}_0 - (2\,E\,\overline{Y}_{zx} - \beta_z - \beta_x)(\beta_z + \beta_x) p_{zx} -$$

$$- (2\,E\,(\overline{Y}_z) - \beta_z)\beta_z p_z + (2\,E\,(\overline{Y}_x) - \beta_x)\beta_x p_x + p_{zx} E s_{zx}^2 + p_z E s_z^2 - p_x E s_x^2 -$$

$$- p_0 E s_0^2 = 0. \quad (129)$$

From

$$p_{zx}(124) + p_z(123) - p_x(122) - p_0(121),$$

we conclude

$$\beta_x(p_{zx} - p_x) + \beta_z(p_{zx} + p_z) = p_{zx} E\,\overline{Y}_{zx} + p_z E\,\overline{Y}_z - p_x E\,\overline{Y}_x - p_0 E\,\overline{Y}_0. \quad (130)$$

Knowing all $\overline{Y}$, $s$, and $p$-values we obtain from (129) and (130) an estimator of $\beta_x$:

$$\frac{p_z p_{zx}(\overline{Y}_{zx} - \overline{Y}_z) - p_0 p_x(\overline{Y}_x - \overline{Y}_0)}{p_z p_{zx} - p_0 p_x} \pm$$

$$\pm \sqrt{\frac{(\overline{Y}_{zx} - \overline{Y}_z - \overline{Y}_x + \overline{Y}_0)^2\, p_0 p_x p_z p_{zx} + (p_z p_{zx} - p_0 p_x)(p_0 s_0^2 + p_x s_x^2 - p_z s_z^2 - p_{zx} s_{zx}^2)}{(p_z p_{zx} - p_0 p_x)^2}}$$

In this way we obtain a min. and a max. value for $\beta_x$. Observe that the estimator is uncertain if $p_z p_{zx} - p_0 p_x$ is small. If, e.g. $z$ has no effect on $x$, $p_0 = p_z$, $p_x = p_{zx}$ and $p_z p_{zx} - p_0 p_x = 0$, i.e. only if $z$ has a (considerable) effect on $x$ can we obtain a usable estimator of $\beta_x$.

If $z = 0$ is a sufficient condition for $x = 0$ (cf. model IV), i.e. $p_0 = 1$ and $p_x = 0$, the estimator of $\beta_x$ can be simplified:

$$\overline{Y}_{zx} - \overline{Y}_z \pm \sqrt{\frac{s_0^2 - p_z s_z^2 - p_{zx} s_{zx}^2}{p_z p_{zx}}}.$$

If $s_0^2 = s_z^2 = s_{zx}^2$ we get the further simplification: $\overline{Y}_{zx} - \overline{Y}_z$.

CHAPTER 8

# CAUSAL DIRECTION PROBLEMS

## 8.0. Introduction

The problem of causal direction has been touched upon in section 1.2, and in the examples (e) and (f) in 1.24. In this chapter some aspects of the statistical treatment of this type of problem are presented. The approaches presented here are related to the instrumental variable approaches in Chaps. 5–7.

## 8.1. A new approach

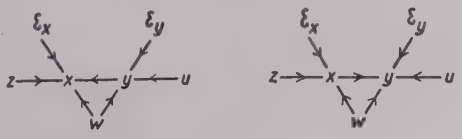
We start from the situation where the presence of an aggregated effect can be suspected (cf. examples in 1.24 and 2.33).

We divide the material into two sub-populations as in example (e). For sub-population I it holds that $y$ has an effect on $x$ and for sub-population II that $x$ has an effect on $y$. The arrow schemes are $y \rightarrow x$ and $x \rightarrow y$, respectively. We denote the number of individuals belonging to sub-population I by $N'$, and those belonging to sub-population II by $N''$. $N'$ and $N''$ are unknown.

*Aim.* We want to decide on the basis of observation material if one of the arrows can be ignored. The arrow $y \rightarrow x$ can be ignored if $y$ has no effect on $x$ or if $N' = 0$. The arrow $x \rightarrow y$ can be ignored if $x$ has no effect on $y$ or if $N'' = 0$.

*Assumption:* We assume that there exists one instrumental variable $z$ which has an effect on $x$ but is independent of $y$ and $w$. We further assume that there exists another instrumental variable, $u$, which has an effect on $y$ but is independent of $x$, $w$, and $z$. The relations between the variables are assumed to be linear. Note that a factor $w$ has been added. We assume that the effects of $w$, $z$, and $u$ on $x$ are the same in the two sub-populations. Analogically, we assume that the effects of $w$, $z$ and $u$ on $y$ are the same in the two sub-populations $\varepsilon_x$ and $\varepsilon_y$ are residual variables.

*Arrow schemes:*




I. II.

*Notation:* By $x'$ and $y'$ we symbolize the values of $x$ and $y$ in sub-population I and by $x''$ and $y''$ the values in sub-population II. By $X$ and $Y$ we denote the observed values of $x$ and $y$.

*Observations:* We have observed $X$, $Y$, $z$, and $u$ for the whole population.

*Approach:* We can choose $Ez = 0$, $Eu = 0$, $Ew = 0$, $Ey' = 0$, $Ex' = 0$, $Ey'' = 0$, $Ex'' = 0$. For sub-population I it holds

$$x' = b'_{xz}\, z' + b'_{xw}\, w' + b'_{xy}\, y' + \varepsilon'_x,$$
$$y' = b'_{yu}\, u' + b'_{yw}\, w' + \varepsilon'_y.$$

For sub-population II it holds

$$x'' = b''_{xz}\, z'' + b''_{xw}\, w'' + \varepsilon''_x,$$
$$y'' = b''_{yu}\, u'' + b''_{yw}\, w'' + b''_{yx}\, x'' + \varepsilon''_y.$$

Since we assume that the effects of $z$, $w$, and $y$ are the same in the two sub-populations:

$$b'_{xz} = b''_{xz},\quad b'_{yu} = b''_{yu},\quad b'_{xw} = b''_{xw},\quad b'_{yw} = b''_{yw}. \tag{131}$$

After elimination of the superscripts on the $b$-, $z$-, $u$- and $w$-values the equations above are reduced to:

$$\left.\begin{aligned} x' &= b_{xz}\, z + b_{xw}\, w + b_{xy}\, y' + \varepsilon'_x \\ y' &= b_{yu}\, u + b_{yw}\, w + \varepsilon'_y \\ x'' &= b_{xz}\, z + b_{xw}\, w + \varepsilon''_x \\ y'' &= b_{yu}\, u + b_{yw}\, w + b_{yx}\, x'' + \varepsilon''_y \end{aligned}\right\} \tag{132}$$

We have assumed (cf. the arrow scheme) that $w$, $z$, and $u$ are independent. Hence it holds that

$$Ewz = 0, \text{ and } Ewu = 0.$$

Now after some simplifications we obtain the following covariances:

$$Ex'z = b_{xz}\, Ez^2 \text{ and } Ex''z = b_{xz}\, Ez^2, \tag{133}$$

$$Ey'z = 0 \text{ and } Ey''z = b_{yx}\, b_{xz}\, Ez^2, \tag{134}$$

$$Ex'u = b_{xy}\, b_{yu}\, Eu^2 \text{ and } Ex''u = 0, \tag{135}$$

$$Ey'u = b_{yu}\, Eu^2 \text{ and } Ey''u = b_{yu}\, Eu^2. \tag{136}$$

Further it holds that

$$E\left(\frac{x'N' + x''N''}{N' + N''}\right) = 0, \quad E\left(\frac{y'N' + y''N''}{N' + N''}\right) = 0. \tag{137}$$

By means of (133) and (136) we obtain $b_{xz}$ and $b_{yu}$ respectively:

$$b_{xz} = \frac{EXz}{Ez^2}, \quad b_{yu} = \frac{EYu}{Eu^2}. \tag{138}$$

By means of (134) and (137) we obtain

$$\frac{N'' b_{yx} b_{xz} Ez^2}{N' + N''} = EYz. \tag{139}$$

By means of (135) and (137) we obtain

$$\frac{N' b_{xy} b_{yu} Eu^2}{N' + N''} = EXu. \tag{140}$$

If $EXu = 0$, i.e. if no correlation exists between $x$ and $u$, we conclude from eq. (140) that either $N' = 0$ or $b_{xy} = 0$ or $b_{yu} = 0$. If $b_{yu} \neq 0$ (cf. eq. 138), i.e. if a correlation exists between $y$ and $u$, the only solution remaining is $N' = 0$ or $b_{xy} = 0$. If $N' = 0$ or $b_{xy} = 0$ we can eliminate the arrow $y \to x$. In the same way we can prove that if

$$EYz = 0 \text{ and if } b_{xz} \neq 0$$

either $N' = 0$ or $b_{xy} = 0$ and hence the arrow $x \to y$ can be eliminated.

Note that the *average effect* of $x$ on $y$ can be written $(N' \cdot 0 + N'' b_{yx})/(N' + N'')$, since the $x$-effect on $y$ for sub-population I is zero and for sub-population II is $b_{yx}$.

According to eqs. (138) and (139) we then obtain $\frac{EYz}{EXz} = \frac{N'' b_{yx}}{N' + N''}$; i.e. $\frac{s_y r_{yz}}{s_x r_{xz}}$ is an estimator of the $x$-effect on $y$. This estimator is the same as that given by Reiersøl (see 6.37).

## 8.2. References

The method used here is related to that given by Working, 1927. The use of exogenous variables, given for instance in articles in Cowles Commission monograph 14 (Hood and Koopmans, ed., 1953), is also related. Simon (1953, 1954) discusses how the "causal ordering" between observed variables can be determined by using a third variable, $z$. The division of the material into two sub-populations is the principal difference between the approach in 8.1 and the methods given by Working and Simon.

APPENDIX

# SOME APPLICATIONS

*Purpose:* In this appendix we intend to illustrate by some examples the approaches and models mentioned in Chaps. 5 and 7.

For these examples the variables and the hypothetical arrow schemes are specified. We discuss the extent to which the assumptions are realistic (plausible). Also certain limitations in the approaches are discussed. Calculations are carried out for only some of the examples. For all examples the effect of $x$ on $y$ is studied.

*Notation:* $x$, $y$, $M$, and $w$ have the same meaning as in Chaps. 5 and 7.

## Ex. for app. I*a*

*References:* Strindberg, 1956.

*Variables:* By $x$ is denoted different classes of root filling:

$x = 0$ Class A = canal was not filled to the apex,
Class B = canal was filled to the apex, or there was slight overfilling,
$x = 1$ Class C = moderate or marked overfilling,
$y = 1$ successful root filling (according to X-ray examination 4 years after the root filling was performed),
$y = 0$ unsuccessful root filling,
$M = 0$ material followed up,
$M = 1$ material not followed up,

*Population:* Strindberg's patients.

*Material:* From Strindberg the following data were obtained:

| Root filling | Material excluded from analysis No. roots | Material followed up No. roots |
|---|---|---|
| Class A | 100 | 226 |
| Class B | 121 | 326 |
| Class C | 65 | 222 |
| Total (roots) | 286 | 774 |

From the table we see that class C is over-represented in the material followed up and class A is under-represented. Every root is an observation unit.

*Arrow scheme:* $x \rightarrow -y \rightarrow -M$.

*Comment:* From the table we see that a large group of patients have not been followed up. It is possible that some of the patients have changed dentists since they suffered from toothache after the root filling was performed and that this is the reason why they are not included in the material followed up.

*Aim:* Estimation of a difference as in app. I *a*.

*Estimation:* In the analysis of the material two other relevent factors have also been considered, (i) number of roots per tooth and (ii) initial state of the periradicular bone. The following data have been calculated (see the table).

TABLE

| No. of roots per tooth | Initial state of periradicular bone ($x$-ray) | A $x=0$ | C $x=1$ | Population (followed-up + excluded) A | C | Estimated differences |
|---|---|---|---|---|---|---|
| 1 | No. rarefaction | $\frac{3}{43}=7\,\%$ | $\frac{16}{55}=29\,\%$ | 62 | 71 | $\left(\frac{16}{71}-\frac{3}{62}\right)\left(\frac{40\cdot 71-39\cdot 62}{16\cdot 40-3\cdot 39}\right)=0.14$ |
| | Rarefaction observed | $\frac{6}{21}=29\,\%$ | $\frac{17}{47}=36\,\%$ | 30 | 61 | $\left(\frac{17}{61}-\frac{6}{30}\right)\left(\frac{15\cdot 61-30\cdot 30}{17\cdot 15-6\cdot 30}\right)=0.01$ |
| 2 | No. rarefaction | $\frac{0}{71}=0\,\%$ | $\frac{6}{35}=17\,\%$ | 102 | 45 | $\left(\frac{6}{45}-\frac{0}{102}\right)\left(\frac{71\cdot 45-29\cdot 102}{6\cdot 71-0\cdot 29}\right)=0.07$ |
| | Rarefaction observed | $\frac{4}{17}=24\,\%$ | $\frac{9}{33}=27\,\%$ | 25 | 43 | $\left(\frac{9}{43}-\frac{4}{25}\right)\left(\frac{13\cdot 43-24\cdot 25}{9\cdot 13-4\cdot 24}\right)=-0.10$ |

The ratios in columns $x=1$ and $x=0$ should be read:

$$\frac{\text{number of successful treatments}}{\text{total number of treatments}} = \text{percentage of successes.}$$

The estimated differences are presented in the last column. (Cf. Table 40, Strindberg, 1955.)

*Discussion:* The estimated differences are uncertain since the population is small.

**Ex. for app. I *a***

*Reference:* Vahlquist-Melander, 1959. (Cf. situation (b) in 4.21).

*Variables:* $x=0$: breast fed from birth $-2\frac{1}{2}$ months,
$x=1$: breast fed from birth $-6\frac{1}{2}$ months or more,
$y=0$: Phosphor in serum less than 6 mg % at the age of 7 months,
$y=1$: Phosphor in serum more than 6 mg % at the age of 7 months.

*Population:* 402 children from northern Sweden.
*Material:*

| $x$ | $y$ | Population | Observed material |
|---|---|---|---|
| 0 | 0 | $N_0$ | $n_0 = 53$ |
| 0 | 1 | $N_y$ | $n_y = 115$ |
| 1 | 0 | $N_x$ | $n_x = 62$ |
| 1 | 1 | $N_{xy}$ | $n_{xy} = 61$ |

We know that $N_0 + N_y = 240$ and $N_x + N_{xy} = 162$.

Without knowing the values of $x$, a large number of values observed for $y$ have been rejected since there seems to be a gross error. The original material includes 402 individuals and the remaining material 291 individuals.

*Arrow scheme:* (Hypothetical): $x \rightarrow -y \rightarrow -M$.

*Estimation:* We want to estimate a difference (cf. eq. (9)).

*Discussion:* It can be doubted that there really is no arrow between $x$ and $M$. Phosphor concentration in serum is a continuous variable but we have constructed a dichotomous variable $y$. For the original continuous variable ($Y$) it probably holds that $x \rightarrow - Y \rightarrow - M$, but for the new variable $y$ it perhaps holds that

## Ex. for app. I *b*

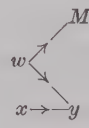
See the example for situation (c) in 4.21 and 3.61 $\gamma$. Moreover note that the missing-plot technique (cf. 3.42) corresponds to this approach.

## Ex. for app. I *c*

*Variables:* $x = 0$: girls, $x = 1$: boys.
$y = 0$: not mature enough to begin at school according to a test,
$y = 1$: mature enough to begin at school according to a test,
$w_1$: number of brothers and sisters, $w_2$: attributes of the mother.
$w_1$ and $w_2$ are $w$-factors, denoted by $w$ in the arrow scheme below.

*Population:* Children born in 1952 in a certain town. The test was performed in 1958. Children who have moved away were not tested.

*Arrow scheme:*



*Comment:* Sex and number of brothers and sisters are independent. Big families have perhaps moved away to a larger extent than have small families (or perhaps vice versa).

*Material:* $x$ and $y$ are known only for the tested children.

*Aim:* Estimation of the difference given in app. I $c$.

### Ex. for approaches I $d$, I $e$, I $f$, and I $g$

Examples for these approaches are not given here but can be constructed by adding residual factors $w_1$ or $w_2$ to the examples given for app. I $a$ and app. I $b$ above.

### Ex. for app. I $h$

*Reference:* Lundström, 1952.

*Variables:* $x = 0$ the mother did not have rubella during the first three months of pregnancy,
$x = 1$ the mother had rubella during the first three months of pregnancy,
$y = 0$ no malformation at birth, $y = 1$ malformation at birth,
$w = 0$ hospital that was cooperative,
$w = 1$ hospital that was not cooperative.

*Population:* Children of certain mothers who had rubella during pregnancy ($x = 1$) and a matched control group (cf. 6.34). Data were reported from the hospitals where the children were born.

*Material:* The material is not complete. Some of the children born in hospitals were not reported perhaps due to absent-mindedness or lack of cooperation. It is also plausible that children suffering from malformation or debility were more often reported than sound children, since physicians are more interested in malformed children. The physicians, who took a real interest in the investigation probably reported every rubella case. From a special investigation concerning pregnant women we know for a considerably large group of women whether or not they had rubella during the first three months of pregnancy. This group was followed up and in this way it was possible to get information concerning the $x$-variable for the missing data. Data for $y$, however, are incomplete.

*Arrow scheme* (hypothetical): See p. 69.

*Aim:* Estimation of the ratio, $\bar{K}$ discussed in app. I $h$.

### Ex. for app. I $k$

Cf. the example from Berkson discussed in 4.23. If for the whole population we had complete data concerning $x$, a solution would be possible as in app. I $k$.

### Ex. for app. II *a* (and app. II *c*)

*Variables:* $x = 0$ individuals not having rheumatism, e.g.
$x = 1$ individuals having rheumatism,
$y = 1$ becoming ill with a contagious disease, $y = 0$ the contrary to $y = 1$,
$z = 0$ not vaccinated against the disease, $z = 1$ vaccinated against the disease,
$M = 0$ response, $M = 1$ not at home (non-response).

*Population:* Individuals in a region. An experiment with a special vaccine was undertaken. The experiment and control groups are included in the population.

*Material:* $x$ and $z$ are known for the whole population. Data concerning $y$ are obtained through interviews.

*Arrow scheme:*

*Comment:* Rheumatic and diseased people are more often at home.
*Aim:* Estimation of a difference as in app. II *a*. See p. 77.

### Ex. for app. II *a*

*Variables:* $x = 0$ non-urban, $x = 1$ urban,
$y = 0$ non-infected, $y = 1$ infected,
$z = 0$ immune, because they have earlier had the same infection,
$z = 1$ non-immune,
$M = 0$ diagnosis clear; i.e. it is known if $y = 0$ or $y = 1$,
$M = 1$ diagnosis unclear.

*Population:* People in urban and non-urban regions.

*Arrow scheme:*

*Comment:* It is often unclear whether or not the infection is present. Diagnoses among the urban population are often more correct due to better medical examinations.

*Material:* $x$ and $z$ are known. Data concerning infection are collected for every individual in the population.

Note: $N_{zy} = 0$ and $N_{xzy} = 0$ because immunity implies $y = 0$.

*Aim:* Estimation of the difference mentioned in app. II *b*. We prefer to estimate just this difference since $x$ has an effect on $z$.

### Ex. for app. II *b* (and app. II *d*)

*Variables:* See the above example; however, we assume now that

$z = 0$ not dead, $z = 1$ dead.

The infection is known to be very severe.

*Arrow scheme:* See p. 77.

*Population, observations and aim* as in the above example.

### Ex. for app. IV

*Reference:* Vahlquist-Melander, 1959.

*Variables:* The $y$-variable in the example on p. 135 is assumed to be measured in mg % (and is not dichotomous); the $x$-variable is unchanged.

*Population, observations and arrow scheme* as in this example.

### Ex. for app. IV

*Reference:* Lundström, 1952.

*Variable:* The $y$-variable in the example on p. 137 is substituted by birth weight.

*Arrow scheme:* $x \rightarrow -y \rightarrow -M$.

*Assumptions:* $R(y)$ is non-increasing.

### Ex. for app. V

*Reference:* Example for situation (d) given in 4.21.

*Arrow scheme:* See p. 75.

*Variables, population and material* are defined in the example for situation (d) given in 4.21 ($x = 0$, women, $x = 1$, men).

*Assumptions:* $R_x(y) \leqslant R_0(y)$

$R_x(y)$ and $R_0(y)$ are non-increasing.

### Ex. for app. V

*Variables:* $x = 0$ worker, $x = 1$ middle class,
$y$ number of teeth intact,
$M = 0$ does not wear dentures, $M = 1$ wearer of dentures.

*Population:* People, 40–60 years old, living in a region.

*Assumptions:* $R_x(y)$ and $R_0(y)$ are non-increasing.

### Ex. for model III in Chap. 7

*Reference:* Lundström, 1952.

*Variables:* $x=1$ the mother had rubella during the first three months of pregnancy
$x=0$ the mother did not have rubella during the first three months of pregnancy
$y$ birth weight
$w=2$ the mother strongly exposed to infection
$w=1$ the mother exposed to various infections
$w=0$ the mother not exposed to infection
$z=0$ the mother had rubella before 1951
$z=1$ the mother did not have rubella before 1951.

*Population:* See Ex. for app. I *h*, p. 137. The material was collected during a rubella epidemic in 1951.

*Material:* $x$, $y$, and $z$ are known for every observation unit (mother + child). Let us ignore the missing data problem and concentrate our interest on the bias due to a common causal factor.

*Aim:* To study the effect of $x$ on $y$.

*Arrow scheme:* See model III, p. 111. It is plausible that $w$ has an effect on both $x$ and $y$ and it is posssible that a virus-infection may be a cause of low birth weight. Hence mothers strongly exposed to various infections may often have extremely small children. Moreover such mothers run a larger risk of contracting rubella. Hence $w$ is a common causal factor.

Mothers who have already had rubella ($z=0$) are immune and there is a strong correlation between $x$ and $z$. Since the preceding rubella epidemic in Sweden occurred more than ten years ago it is reasonable to assume that the unknown correlation between $z$ and $w$ is about zero. At least if the material is divided with respect to mother's age and living place (urban or not), the correlation must be unimportant. Moreover it seems reasonable to assume that $z$ has no other effect on $y$ than via $x$. Model III is then applicable.

### Ex. for model IV

*Variables*, *population*, *material* and *aim* as in the Ex. for model III.

*Arrow scheme:* See model IV (p. 112). This arrow scheme differs from the preceding one in one important respect: $z$ is assumed to have a direct effect on $y$. This assumption is probably not necessary.

Since there are almost no observations for which $z=0$ and $x=1$ only three combinations of $x$- and $z$-values are possible, as in model IV.

## Ex. for model VI

*Reference:* Westin-Wold (1943): *1942 års tandmönstring av inskrivningsskyldiga,* and Welander (1955).

*Variables:* $x = 0$ no tartar on the teeth, $x = 1$ tartar on the teeth
$y$ caries status (number of decayed surfaces)
$w$ factors independent of $z$ and $t$
$z = 1, t = 1$ bad oral hygien
$z = 0, t = 1$ fairly good oral hygiene
$z = 0, t = 0$ good oral hygien.

*Population*: 6468 Swedish recruits, 20 years old.

*Material:* Data concerning $x$, $y$, $z$, and $t$. For the individuals only four combination of values for $x$, $z$, and $t$ are represented: i) $x = 0$, $z = 0$, $t = 0$, ii) $x = 0$, $z = 1$, $t = 0$, iii) $x = 1$, $z = 1$, $t = 0$, iv) $x = 1$, $z = 1$, $t = 1$.

*Arrow scheme:* (See p. 118.) The $w$-factor may e.g. be special yeast in the oral cavity.

*Aim:* To study if there exists a correlation between $x$ and $y$ that cannot be explained by $z, t,$ or $w$. The assumption concerning interactions on p. 121 is reasonable, but it is difficult to prove anything definitely.

# REFERENCES

## *Abbreviations*

*AMS* Annals of Mathematical Statistics.
*JAMA* Journal of the American Medical Association.
*JASA* Journal of the American Statistical Association.
*JRSS* Journal of the Royal Statistical Society.
SOU Statens Offentliga Utredningar (Stockholm).


Abramson, E. *et al.* (1954), Fluor som medel mot tandröta. *Svensk Tandläkartidskrift*, Vol. 47, Nr. 1, 1954.

Bellman, R. (1957), *Dynamic Programming*, Princeton, NJ.

Berkson, J. (1955), The statistical study of association between smoking and lung cancer. *Proceedings of the Staff Meetings of the Mayo Clinic*, Vol. 30, 1955.

Berkson, J., English, J. P. and Willius, F. A. (1940), Tobacco and coronary disease. *JAMA*, 1940.

Birnbaum, Z. W. and Sirken, M. G. (1950), On the total error due to non-interview and to random sampling. *International Journal of Opinion and Attitude Research*, 1950.

Bradford-Hill, A. (1956), *Principles of medical Statistics*. 6th edition. London.

Carlson, Gösta (1954), Statistik och orsaksforskning. *Sociala Meddelanden* nr. 9, 1954.

Chapin, F. S. (1947), *Experimental Designs in Sociological Research*. New York.

Cochran, W. G. (1951), Modern methods in the sampling of human populations. *Amer. Journal of Public Health*, 1951.

—— (1954), Some methods for strengthening the $\chi^2$-test. J. of *Biometrics*.

Cornfield, J. (1942), On certain biases in sampling of human populations. *JASA*, 1942.

Cramér, H. (1945), *Mathematical Methods of Statistics*. Uppsala.

Dalenius, T. (1955), The problem of not at homes. *Statistical Review nr. 4*, 1955.

—— (1957), *Sampling in Sweden*. Thesis. Uppsala.

—— (1957), Bortfallsproblemet vid statistiska undersökningar. *Den Svenska Marknaden*, 1957.

Durbin, J. (1954), Errors in variables. *Rev. Inst. Int. Stat.* Vol. 22, 1954.

—— (1954), Non-response and call-backs in surveys. *Bull. of the Int. Stat. Inst.* Tome XXXIV, 2nd Part, Rome, 1954.

Ezekiel, M. (1941), *Methods of correlation analysis*. 2nd edition. New York.

Fisher, R. A. (1935), *The design of experiments*. Edinburgh and London.

Frisch, R. (1934), Statistical confluence analysis by means of complete regression systems. *University Institute of Economics*, Oslo, Publication no. 5.

Gallup, G. (1938), Government and the sampling referendum. *JASA*, 1938, Vol. 33, no. 201.

Geary, R. C. (1949), Studies in relations between economic time series. *Econometrica*, Vol. 17.

—— (1953), Non-linear functional relationship between two variables when one variable is controlled. *JASA*, 1953.

Gesell, A. (1942), The method of co-twin control. *Science*, Vol. 95, 1942.

GUMBEL, E. J. (1958), *Statistics of Extremes*. New York.
HALD, A. (1946), Den afstumpede, normale Fordeling. *Matematisk Tidsskrift, B*, 1946.
HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953), *Theory and Practice of Sampling*. Vol. 1. New York.
HANSEN, M. H. and HURWITZ, W. N. (1946), The problem of non-response in sample surveys. *JASA*, 1946.
HENDRICKS, W. A. (1949), Adjustment for bias by non-response in mailed surveys. *Agr. Ec. Res.*, 1949.
v. HOFSTEN, E. (1948), *Praktisk statistik*. Stockholm.
HYRENIUS, H. (1945), Befolkningsrörelsens beroende av konjunkturerna under tiden efter 1920. *SOU* 1945: 53.
HUSÉN, T. (1953), *Tvillingstudier*. Stockholm.
KENDALL, M. G. (1948), *The Advanced Theory of Statistics*. Vol. II. 2nd edition. London.
KEMPTHORNE, O. (1955), The randomization theory of experimental inference. *JASA*, 1955.
KEMPTHORNE, O, editor (1954), *Statistics and Mathematics in Biology*. Ames, Iowa, 1954.
—— (1957), *An Introduction to Genetic Statistics*. New York.
KIMBALL, A. W. (1957), *Disease incidence estimation in populations subject to multiple causes of death*. Uppsala. Reprint.
LI C. C. (1956), The concept of path coefficient and its impact on population genetics. *Biometrics 12*.
LUNDBERG, G. A. (1948), *Social Research. A study in methods of gathering data*. New York.
LUNDSTRÖM, R. (1952) Rubella during pregnancy. *Acta Pædiatrica*.
MALINVAUD, E. (1954), *Réflexions sur l'emploi des modèls a erreurs sur les variables*. Reprint.
MALMQUIST, S. (1948), *A statistical analysis of the demand for liquor in Sweden*. Thesis. Uppsala.
MAYO-SMITH, R. (1895), *Statistics and Sociology*. New York. Reprint 1910.
MINER, J. R. (1945), Some uses of statistical methods in medicine. *J. of Biometrics*, 1945.
MINOT, C. S. (1891), Senescence and rejuvenation. *J. Physiol.*, 12.
MOBERG, A. (1951), *Vem blev student och vad blev studenten*. Thesis. Lund.
NEWMAN, H. H., FREEMAN, F. N. and HOLZINGER, K. J. (1936), *Twins: A study of heredity and environment*. Chicago, 1936.
NEYMAN, J. (1934), On the two different aspects of the representative method. The method of stratified sampling and the method of purposive selection. *JRSS*, 97, 1934.
NYQUIST, G. (1952), *A Study of Denture Sore Mouth*. Thesis. Uppsala.
POLITZ, A. and SIMMONS, W. (1949), (1950), An attempt to get the "Not at Homes" into the sample without callbacks. *JASA*, 1949 and 1950.
REIERSØL, O. (1941), Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica*, 9, 1941.
—— (1945), *Confluence Analysis by means of Instrumental Sets of Variables*. Thesis. Stockholm, 1945.
RUSSELL, B. (1929), On the notion of cause, with application to the free-will problem. In: *Our knowledge of the external world*. London.

Simon, H. A. (1953), Causal ordering and identifiability. In: W. C. Hood and T. C. Koopmans, eds. (1953): *Studies in Econometric Method*. New York.
—— (1954), Spurious correlation: a causal interpretation. Cowles Comm. Papers. New Series, No. 89. *JASA*, Vol. 49, 1954.
—— H. A. (1957), *Models of Man*. New York.
Strindberg, L. Z. (1956), *The Dependence of the Results of Pulp Therapy on certain Factors*. Thesis. Stockholm.
Tinbergen, J. (1939), *Statistical testing of business cycle theories, II. Business cycles in the United States of America 1919–32*. Geneva.
Tukey, J. W. (1954), Causation, regression, and path analysis. *Statistics and Mathematics in Biology*. Ames, Iowa, 1954.
Vahlquist, B. and Melander, O. (1959), Breast feeding and artificial feeding. *Acta Pædiatrica*. (in press).
Wallis, W. A. and Roberts, H. V. (1956), *Statistics. A new approach*. Illinois.
Welander, E. (1955), *The Occurence of dental Caries in the Permanent Dentition*. Thesis. Uppsala.
Westin-Wold, (1943): *1942 års tandmönstring av inskrivningsskyldiga*, Odont. Tidskr. fasc. 5.
Wilson, E. B. (1952), *An Introduction to Scientific Research*. New York.
Wold, H. O. (1940), Efterfrågan på jordbruksprodukter och dess känslighet för pris- och inkomstförändringar. *SOU* 1940:16.
Wold, H. O. and Juréen, L. (1952), *Demand Analysis*. Uppsala.
Wold, H. O. (1954), Causality and econometrics. *Econometrica*, 1954.
—— (1956), Causal inference from observational data. *JRSS*, 1956.
Wold, H. O. and Faxér, P. (1957), On the specification error in regression analysis. *AMS*, Vol. 28, No. 1.
Wold, H. O. (1958), On the definition of causation and related concepts. (Manuscript).
Working, H. (1927), What do statistical demand curves show. *Quart. J. of Econ.*, 41.
Wright, S. (1918), On the nature of size factors. *Genetics*, 3.
—— (1920), The relative importance of heredity and environment in determining the piebald pattern of quinea pigs. *Proc. Nat. Acad. Sci.*, 6, 1920.
—— (1921), Correlation and causation. *J. Agr. Res.*, 20.
—— (1934), The method of path coefficients. *AMS*, 5. 1934.
—— (1954), The interpretation of multivariate systems. *Statistics and Mathematics in Biology*. Ames, Iowa, 1954.
Yates, F. (1933), Principles of orthogonality and confounding in replicated experiments. *J. Agr. Sci.*, 23.
—— (1938), The comparative advantage of systematic and randomized arrangements in the design of agricultural and biological experiments. *Biometrika*, Vol. 30, 1938.
Yule, G. U. (1907), On the theory of correlation for any number of variables treated by a new system of notation, *Proceeding of the Royal Society*.
—— (1911), *An Introduction to the Theory of Statistics*. London.
—— (1926), Why do we sometimes get nonsense correlations between time-series? *JRSS*, Vol. 89, 1926.
Yule, G. U. and Kendall, M. G. (1937), *An Introduction to the Theory of Statistics*. London.